

Annamaria De Santis and Irene Rossi (Eds.)

**Crossing Experiences in Digital Epigraphy. From Practice to Discipline**



Annamaria De Santis and Irene Rossi (Eds.)

# **Crossing Experiences in Digital Epigraphy**



From Practice to Discipline

Managing Editor: Katarzyna Michalak

Associate Editors: Francesca Corazza  
and Łukasz Połczyński

Language Editor: Rebecca Crozier

**DE GRUYTER**

ISBN 978-3-11-060719-2  
e-ISBN 978-3-11-060720-8



This work is licensed under the Creative Commons Attribution 4.0 International license (CC BY 4.0)  
For details go to <http://creativecommons.org/licenses/by/4.0>

© 2018 Annamaria De Santis, Irene Rossi and chapters' contributors

Published by De Gruyter Poland Ltd, Warsaw/Berlin

Part of Walter de Gruyter GmbH, Berlin/Boston

The book is published with open access at [www.degruyter.com](http://www.degruyter.com).

**Library of Congress Cataloging-in-Publication Data**

A CIP catalogue record for this book has been applied for at the Library of Congress.

Managing Editor: Katarzyna Michalak

Associate Editors: Francesca Corazza and Łukasz Połczyński

Language Editor: Rebecca Crozier

[www.degruyter.com](http://www.degruyter.com)

Cover illustration: Łukasz Połczyński; Ancient South Arabian inscription (Ṣan'ā', Military Museum, MŞM 149)

# Contents

## Introduction — XIII

- The Experience of DASI Project — XIV
- Concept and Content of the Volume — XV
- Reading Path — XVII
- Acknowledgements — XVIII

Alessandra Avanzini, Annamaria De Santis and Irene Rossi

## 1 **Encoding, Interoperability, Lexicography: Digital Epigraphy Through the Lens of DASI Experience — 1**

- 1.1 Digitizing the Epigraphic Heritage of Ancient Arabia: From CSAI to DASI — 1
- 1.2 Data Modelling and Textual Encoding — 3
  - 1.2.1 The Data Model: XML vs Database — 3
  - 1.2.2 The Conceptual Model: Text vs Object — 5
  - 1.2.3 Encoding for Curated Digital Editions: In-Line vs External Apparatus Criticus — 7
- 1.3 Interoperability — 9
  - 1.3.1 Text Encoding and Representation: Standards vs Specificities — 9
  - 1.3.2 Harmonization of Metadata — 10
  - 1.3.3 Openness and Semantic Interoperability — 12
- 1.4 Lexicography — 13
  - 1.4.1 Approach to Under-Resourced Languages — 13
  - 1.4.2 Translations — 15
- 1.5 Conclusions and General Remarks — 16
  - Bibliography — 16

## Part I: Data Modelling and Encoding for Curated Editions and Linguistic Study

Christiane Zimmermann, Kerstin Kazzazi and Jens-Uwe Bahr

## 2 **Methodological, Structural and Technical Challenges of a German-English Runic/*RuneS* Database — 21**

- 2.1 Introduction — 21
  - 2.1.1 The Main Research Areas and the Specific Profile of *RuneS* — 21
  - 2.1.2 *RuneS* and Digital Epigraphy — 22
  - 2.1.3 Why is a Digital *RuneS* Database Necessary? — 23
- 2.2 Design of the Database — 24
  - 2.2.1 Design of the Database – Step Zero: Basic Considerations — 24

- 2.2.2 Design of the Database – Step One: Type of Data? — **24**
- 2.2.2.1 Backbone of the Database: The *Find* Fields — **26**
- 2.2.3 Design of the Database – Step Two: The Graphemic Section and the Structure of the Database — **28**
- 2.2.4 Design of the Database – Step Three: The Bilingual Layout — **31**
- 2.2.4.1 Bilingual Terminology: Choices — **31**
- 2.2.4.2 Bilingual Terminology: Technical Aspects — **32**
- 2.2.5 Design of the Database – Step Four: Data Mask for the Input of Graphic and Graphemic Data — **33**
- 2.3 Concluding Remarks — **34**
- Bibliography — **35**

María José Estarán, Francisco Beltrán, Eduardo Orduña and Joaquín Gorrochategui

- 3 Hesperia, a Database for Palaeohispanic Languages; and AELAW, a Database for the Ancient European Languages and Writings. Challenges, Solutions, Prospects — 36**
- 3.1 Introduction to BDHesp and AELAW Databases — **37**
- 3.2 Palaeohispanic Languages and Writings — **38**
- 3.3 BDHesp (Banco de Datos de Lenguas Paleohispánicas Hesperia) — **40**
- 3.3.1 Developing BDHesp: From an Epigraphic Database to a Databank of Palaeohispanic Languages — **41**
- 3.3.2 Challenges Arising from the Digitalization of Palaeohispanic Epigraphy and Solutions Addressed in BDHesp — **42**
- 3.4 AELAW — **45**
- 3.4.1 Developing of the AELAW Database — **46**
- 3.4.2 Challenges Arising from the Digitalization of Palaeo-European Epigraphy and Solutions Addressed in AELAW — **47**
- Bibliography — **48**

Francesco Di Filippo

- 4 Sinleqiunnini: Designing an Annotated Text Collection for Logo-Syllabic Writing Systems — 49**
- 4.1 The Project — **49**
- 4.2 Collection Design: Mark-Up Languages Versus Database Model — **51**
- 4.3 Sinleqiunnini Data Container — **57**
- 4.4 Conclusions — **61**
- Bibliography — **63**

Christian Prager, Nikolai Grube, Maximilian Brodhun, Katja Diederichs, Franziska Diehr, Sven Gronemeyer and Elisabeth Wagner

**5 The Digital Exploration of Maya Hieroglyphic Writing and Language — 65**

- 5.1 Introduction — 65
- 5.2 Maya Hieroglyphic Writing — 67
  - 5.2.1 Decipherment — 71
  - 5.2.2 Sign Lists and Classification — 72
- 5.3 Digital Epigraphy of Classic Mayan — 73
  - 5.3.1 Documentation of Object Information — 73
    - 5.3.1.1 Controlled Vocabularies — 74
    - 5.3.1.2 Technical Infrastructure — 75
  - 5.3.2 Documentation of Signs and Graphs — 76
    - 5.3.2.1 Modelling Graph Variants — 77
    - 5.3.2.2 Modelling Multiple Sign Functions — 77
    - 5.3.2.3 Evaluating Sign Readings — 78
    - 5.3.2.4 Components for Generating a Digital Corpus — 79
    - 5.3.2.5 A TEI Schema for Digitally Documenting Maya Inscriptions — 80
    - 5.3.2.6 Multi-Level, Semi-Automatic Annotation of Classic Mayan — 80
- 5.4 Summary and Conclusion — 81
  - Bibliography — 82

Alessandro Bausi and Pietro M. Liuzzo

**6 Inscriptions from Ethiopia. Encoding Inscriptions in Beta Maṣāḥəft — 84**

- 6.1 Ethiopian and Eritrean Ancient Epigraphy — 84
- 6.2 Beta Maṣāḥəft — 87
- 6.3 Inscriptions in Beta Maṣāḥəft — 88
  - 6.3.1 The Challenges of Encoding Inscriptions in Semitic Scripts — 88
  - 6.3.2 Multilingual Inscriptions — 90
  - 6.3.3 Inscriptions in Greek — 91
- 6.4 Conclusions — 92
  - Bibliography — 92

Paolo Xella and José Á. Zamora

**7 Phoenician Digital Epigraphy: CIP Project, the State of the Art — 93**

- 7.1 Motive of the Project and Institutional Background — 93
- 7.2 Aims and General Description of the Project — 94
- 7.3 Basic Technical Data — 95
- 7.4 Organization and Structure of the Corpus — 97
- 7.5 State of the Database and Future Outlook — 100
  - Bibliography — 101

Daniel Burt, Ahmad Al-Jallad and Michael C.A. Macdonald

- 8 The Online Corpus of the Inscriptions of Ancient North Arabia — 102**
- 8.1 The Background to OCIANA — 102
- 8.1.1 Building a Digital Corpus: Challenges, Objectives and Perspectives — 106
- 8.2 The Development of OCIANA — 108
- 8.3 The Future of OCIANA — 115
- Bibliography — 116

Anne Multhoff

- 9 A Methodological Framework for the Epigraphic South Arabian Lexicography. The Case of the Sabaic Online Dictionary — 118**
- 9.1 Introduction — 118
- 9.1.1 General Remarks — 118
- 9.1.2 Scope of the Project — 119
- 9.2 Material Base — 120
- 9.2.1 Character of Material — 120
- 9.2.2 Collection of Material — 121
- 9.2.3 Organisation of Material — 121
- 9.3 Morphological Analysis — 122
- 9.4 Definition of Lemmata — 123
- 9.4.1 Treatment of Homographs — 123
- 9.4.2 Deliberate Splitting of Lexemes — 124
- 9.4.3 Heterographs with Identical Meaning — 125
- 9.4.4 Treatment of Incorrect Forms — 125
- 9.5 Presentation of Material — 126
- 9.5.1 Structure of Presentation — 126
- 9.5.2 Accessible Material — 127
- 9.5.2.1 Translation — 127
- 9.5.2.2 Existing Translations — 128
- 9.5.2.3 Etymological Parallels — 129
- 9.5.2.4 Morphological Catalogue — 129
- 9.5.2.5 Examples in Context — 129
- 9.6 Results Reached Thus Far — 130
- Bibliography — 131

Ronald Ruzicka

- 10 KALAM: A Word Analyzer for Sabaic — 133**
- 10.1 An Automatic Word Analyzer for Languages Epigraphically Attested — 133



- 10.2 Requirements of the Word Analyzer for Sabaic — **135**
- 10.3 Functioning of the Word Analyzer — **136**
- 10.3.1 Using KALAM — **137**
- 10.4 Future Perspectives — **139**
- Bibliography — **140**

Jamie Novotny and Karen Radner

- 11 Official Inscriptions of the Middle East in Antiquity: Online Text Corpora and Map Interface — 141**
- 11.1 Introduction — **141**
- 11.2 Overview of OIMEA and Its Sub-Projects — **143**
- 11.2.1 Royal Inscriptions of Assyria Online — **144**
- 11.2.2 Royal Inscriptions of Babylonia Online — **145**
- 11.3 The Map Interface Ancient Records of Middle Eastern Polities — **147**
- 11.4 Methodological Problems and Technical Issues — **150**
- 11.5 Future Prospects — **152**
- Bibliography — **153**

Sébastien Biston-Moulin and Christophe Thiers

- 12 The Karnak Project: A Comprehensive Edition of the Largest Ancient Egyptian Temple — 155**
- 12.1 Introduction — **155**
- 12.2 Towards an Interactive Corpus of Primary Sources in Ancient Egyptian — **157**
- 12.2.1 Fieldwork and Implementation of the Tools — **157**
- 12.2.2 Production and Dissemination of Reference Documents — **159**
- 12.2.3 From Plain Text to Indexed Interactive Text — **161**
- 12.3 Progress and Prospects — **163**
- Bibliography — **164**

## **Part II: Providing Access: Portals, Interoperability and Aggregators**

Gerfrid G.W. Müller and Daniel Schwemer

- 13 Hethitologie-Portal Mainz (HPM). A Digital Infrastructure for Hittitology and Related Fields in Ancient Near Eastern Studies — 167**
- 13.1 Remit and Unique Proposition — **167**
- 13.2 Objectives: Innovation, Collaboration, Acceleration — **169**
- 13.3 History and Status Quo 2017 — **170**

- 13.4 Organization: A Network of Researchers and Projects — **171**
- 13.5 Digital Components and Concepts — **172**
- 13.5.1 Components of HPM — **172**
- 13.5.2 Open Standards and Widespread Open-Source Software — **173**
- 13.5.3 Continuity Online: Development and Experiences — **174**
- 13.5.4 Tools for Scholars, not Scholars for Tools — **175**
- 13.5.5 Connecting Data — **177**
- 13.6 Outlook: Expansion, Connectivity, Sustainability — **178**
- Bibliography — **179**

Nadia Cannata

- 14 EDV – Italian Medieval Epigraphy in the Vernacular Some Editorial Problems Discussed — 180**
- 14.1 The Corpus — **180**
- 14.2 The Background — **181**
- 14.3 History, Geography, Forms and Functions — **182**
- 14.4 How are the Data Organized — **185**
- 14.5 Conclusion — **189**
- Bibliography — **190**

Mark Depauw

- 15 Trismegistos: Optimizing Interoperability for Texts from the Ancient World — 193**
- 15.1 The Development of Trismegistos (Texts) — **193**
- 15.2 New Techniques & Other Trismegistos Databases — **196**
- 15.3 The Raison d’Être of Trismegistos — **198**
- Bibliography — **200**

Adam Rabinowitz, Ryan Shaw and Patrick Golden

- 16 Making up for Lost Time: Digital Epigraphy, Chronology, and the PeriodO Project — 202**
- 16.1 The Promise of Digital Epigraphy — **202**
- 16.2 The Trouble with Time — **204**
- 16.3 The PeriodO Temporal Gazetteer — **206**
- 16.3.1 PeriodO and Digital Epigraphy — **207**
- 16.3.2 Using the PeriodO Gazetteer in Epigraphic Corpora — **209**
- 16.3.2.1 Technical Specifications — **209**
- 16.3.2.2 Reconciliation — **210**
- 16.3.2.3 Adding Data to the Gazetteer — **211**
- 16.3.2.4 EpiDoc Guidelines — **212**
- 16.4 Conclusions — **212**
- Bibliography — **214**

Pietro M. Liuzzo

- 17 EAGLE Continued: IDEA. The International Digital Epigraphy Association — 216**
- 17.1 The EAGLE Project Steps — 216
- 17.1.1 The EAGLE Aggregator — 216
- 17.1.2 The EAGLE Portal — 217
- 17.2 IDEA — 218
- 17.3 Methodological Issues Faced During EAGLE — 219
- 17.4 Methodological Issues Faced After EAGLE — 223
- 17.5 General Issues in Digital Epigraphy — 225
- 17.6 Conclusions — 228
- Bibliography — 228

Thomas Kollatz

- 18 EPIDAT – Research Platform for Jewish Epigraphy — 231**
- 18.1 Introduction — 231
- 18.2 EPIDAT Metadata Collections — 232
- 18.3 Text Encoding — 233
- 18.4 Reuse of Data — 235
- 18.5 Interoperability — 236
- Bibliography — 238

Jonathan R.W. Prag and James Chartrand

- 19 I.Sicily: Building a Digital Corpus of the Inscriptions of Ancient Sicily — 240**
- 19.1 Background — 240
- 19.2 Challenges and Ambitions — 245
- 19.2.1 Text-Editing and Annotation — 245
- 19.2.2 Linked Open Data? — 248
- 19.2.3 Collaboration and Outreach — 249
- 19.3 Conclusions — 251
- Bibliography — 251

**Conclusions — 253**

**Appendix A — 258**

**Appendix B — 289**

**List of Figures and Tables — 293**

**Index — 296**



# Introduction

Epigraphy is a multifaceted discipline. Even more than in manuscript studies or papyrology, a researcher approaching an epigraph should be competent with philology, linguistics, archaeology, history of art, not to speak of history tout-court, being inscriptions studied first of all as primary historical sources. The peculiar nature of the epigraphic document – both textual and physical – has put the reflection on digitization of epigraphs at the crossroads of the discussions and advancements in digital humanities and digital heritage, in addition to computational linguistics.

The digitization of the epigraphic heritage is at an advanced stage. A significant number of projects digitizing inscriptions, of both small and big corpora, with different objectives are either under development, or have been recently completed. Many papers have been written, and several proceedings of meetings and conferences dedicated to this topic have been published.

However, digital epigraphy is not yet considered a proper discipline. Digital epigraphers have acquired their skills in digitization methods and techniques informally, “in the field”, through a progressive refinement of those established in the digital humanities. Scholars interested in digital epigraphy are creating more or less formal networks in order to exchange ideas and suggestions, even in very different historical and geographical domains. Nevertheless, there are still no regular occasions to meet and discuss.

Moreover, this large and across-the-board community does not recognize itself in specific journals. They continue to communicate the results of their scientific and technical activities in journals dealing with traditional epigraphy, or, at best, digital humanities in general.

This book is precisely intended to stimulate debate among those practicing digital epigraphy, by recording the methodological issues they have addressed while carrying out specific projects, the solutions they have applied and the criteria that have led to their choices.

In particular, whereas a consistent number of digital initiatives in the domain of Classical epigraphy have been well represented in the proceedings of conferences organized within the frame of the project EAGLE,<sup>1</sup> other domains – and that of Semitic epigraphy *in primis* – are in a quite different situation. Barriers due to the extreme wealth, and also diversity, of writing systems and languages, and to cultural and historical fragmentation, make confrontation and cooperation difficult.

For this reason, the projects represented in the nineteen contributions collected in this book are intentionally diverse in geographic and chronological context, for script and language, and typology of digital output.

---

<sup>1</sup> See further on in the volume (in particular the contributions by Liuzzo) for detailed bibliography.

## The Experience of DASI Project

The idea of a volume collecting different experiences of projects on digital epigraphy has arisen within the frame of DASI – *Digital Archive for the Study of pre-Islamic Arabian Inscriptions*, an ERC – Advanced Grant funded project led by Prof. Alessandra Avanzini at the University of Pisa, aimed at gathering, in an open-access archive, the curated edition of the epigraphic corpora of pre-Islamic Arabia. These consist of thousands of Ancient South Arabian, Ancient North Arabian and Aramaic inscriptions produced since the beginning of the first millennium BCE until the advent of Islam. The study of these inscriptions is essential in order to fill a significant gap in research on the ancient and late antique Near East.

During the five years of the project, a team (consisting of epigraphers, archaeologists, art-historians, digital humanists and IT specialists) worked together, facing methodological and technological challenges while building upon previous experiences of digitization of inscriptional corpora in Semitic languages and alphabetic scripts.

Basic, common issues concerned the modelling of data in order to best describe the complex nature of the epigraphic source, and the encoding of text for its critical edition. Fundamental issues such as those of compliance to standards, interoperability and data openness were tackled. Moreover, specific methodological and technical challenges were faced when approaching the study of under-resourced languages, such as those of pre-Islamic Arabia, which are documented only by epigraphic sources. Specific, lexicographic tools were designed to enhance the description of the language and thereby reach a better comprehension of the messages conveyed by the inscriptions – ultimately leading to the best possible understanding and dissemination of the history and culture of the peoples inhabiting Arabia in pre-Islamic times.

The DASI project has attempted to make the tradition of studies related to pre-Islamic Arabia less “marginal” than before, making the edition of about 10,000 inscriptions originating from ancient Arabia openly available. It has tried to provide useful tools and suggest new approaches to the study of this rich cultural heritage, and to foster reasoning on best practice by taking account of domain-specific questions. This has led to a constant search for confrontation with other digital epigraphy projects.

This volume, conceived during the post-grant phase of the project, continues the mentioned practice of confrontation, wishing to raise new questions and open further, unexpected research perspectives.

## Concept and Content of the Volume

With this vision in mind, this book gives voice to those who have conceived and carried out diverse projects, ranging: from antiquity to medieval and modern times; from alphabetic to logographic writing systems; from Indo-European to Chamito-Semitic to Ancient American languages; from specific databases and lexica, to aggregators, infrastructures and gazetteers.

Hereafter, summaries of the main characteristics of each project and the topics of the related papers are provided in order to facilitate the readers' orientation.

Chapter 1, by Avanzini, De Santis and Rossi, describes the project DASI – *Digital Archive for the Study of pre-Islamic Arabian Inscriptions*, focusing on the main digital epigraphy themes discussed throughout this volume: text encoding and data modelling, interoperability, and lexicography.

The project RuneS – *Runic writing in the Germanic languages* (Chapter 2) collects texts in different Germanic languages and using different Runic writing systems. This comparative approach to the study of the script has led, as explained in the contribution by Zimmermann, Kazzazi and Bahr, to transcend the existent descriptive systems and enhance the visual documentation of inscriptions, through the tagging of images.

Similarly, *Hesperia – Banco de datos de lenguas paleohispánicas* gathers inscriptions and coins in the different Palaeohispanic languages, written in multiple writing systems. The solutions adopted to register and make searchable both script variants and the different transliterations used in the study tradition, are described by Estarán, Beltrán, Orduña and Gorrochategui in Chapter 3.

The two projects *Sinlequiunnini* (Di Filippo) and *Text Database and Dictionary of Classic Mayan* (Prager, Grube, Brodhun, Diederichs, Diehr, Gronemeyer and Wagner) propose different solutions in the textual data modelling in relation to logo-syllabic writing systems, in particular dealing with languages whose interpretation is highly context-driven, in the first case (Chapter 4), and with a still partially deciphered script, in the second one (Chapter 5).

The *Beta Maṣāḥaft* project (Chapter 6) deals with Ethiopian and Eritrean inscriptions and manuscripts. Bausi and Liuzzo address the issue of encoding in XML the relation among multiple copies of the same epigraphic text in a multilingual context, and of annotating their different scripts.

The CIP – *Corpus Inscriptionum Phoenicarum necnon Poenicarum* (Chapter 7) is the first attempt at carrying out a census of the Phoenician and Punic inscriptions spread in a very wide territory, from the Eastern to the Western Mediterranean. The contribution by Xella and Zamora provides an overview of the criteria they have followed to create a complete edition of the only direct textual sources for the reconstruction of the history and culture of this civilization, in the current absence of any attestation of literary texts.

The OCIANA – *Online Corpus of the Inscriptions of Ancient North Arabia* project (Burt, al-Jallad and Macdonald) is a database mainly designed to catalogue graffiti.

Their curated editions, including transcriptions, transliterations in Latin characters and translations, include encoding with particular attention to grammatical analysis and onomastics (Chapter 8).

As the mentioned projects show, the digitization of the overall epigraphic heritage is often aimed at supporting linguistic study. The *Sabaic Dictionary Online* aims at cataloguing all extant lexical material of one of the Ancient South Arabian languages (Chapter 9). Multhoff provides a sound explanation of the methodological issues concerning the annotation of morphological analysis: treatment of ambiguous forms, homographs, heterographs with identical meaning, variant readings, incorrect forms.

The lemmatizer for the Ancient South Arabian languages, KALAM, performs the automatic detection of morphological attributes (Chapter 10). Ruzicka describes its principles and functioning. The contribution must be considered within the frame of the application of NLP techniques to ancient, under-resourced languages.

The OIMEA – *Official Inscriptions of the Middle East in Antiquity* project (Novotny and Radner) edits all the official inscriptions of ancient Middle Eastern polities in cuneiform script. Texts are geo-referenced and fully lemmatized: lexical and grammatical tagging is carried out in order to create glossaries and allow search of text and translation. Historical research is enhanced by the creation of a map-based interface to access geographical information mentioned in cuneiform sources (Chapter 11).

The project *Karnak* (Biston-Moulin and Thiers) focuses on the epigraphs located *in situ* in the ancient Egyptian temples of Karnak. Therefore particular attention is devoted to the preservation of the relation between the inscriptions and their architectural position. An extensive photographic coverage provides high-resolution orthophotographs flanking the transliterations of hieroglyphic, hieratic and demotic texts. These are the basis for a digital lexicon of the languages documented in the temples (Chapter 12).

The infrastructure of the HPM – *Hethitologie-Portal Mainz* (Chapter 13) provides maintenance and access to several independent digital resources available on Hittitology studies. Müller and Schwemer recall the history of a long-lasting project; the continuous technical updates that have been necessary over time; the specific policies for the attribution of resources, their versioning and intellectual property.

Other projects cope with the establishment of systems to identify, sort and connect digital resources. The interdependence of geographic and chronological entities and their labelling, and the need for ontologies with the objective of structuring this information is exemplified by the project EDV – *Epigraphic Database Vernacular* (Cannata), which collects the vernacular inscriptions produced in Italy from late Medieval to Early Modern Age (Chapter 14).

The *Trismegistos* project (Depauw) aims at implementing an identification system, which attributes an ID to each known ancient inscription. This is a first step to tackle the issue of disambiguating and connecting several editions for the same inscriptions in a LOD environment (Chapter 15).



The objective of the project *PeriodO* (Rabinowitz, Shaw and Golden) is the creation of a Linked Data gazetteer of structured period definitions, which provides links between time periods and geographic and cultural contexts, and translation between absolute dates and relative chronologies. Once applied to digital epigraphy, it will foster interoperability of epigraphic collections and their connection with archaeological datasets (Chapter 16).

Interoperability is fully achieved by the aggregator EAGLE, which collects Greek and Latin epigraphs from many different repositories and makes them available to Europeana. The contribution by Liuzzo focuses on the challenges faced, during and after the end of the project, from the up-conversion to the EAGLE schema of the epigraphic records to the harmonization of the terminologies involved (Chapter 17).

Finally, the EPIDAT – *Database of Jewish Epigraphy* project (Kollatz; Chapter 18), which provides its records to national and European aggregators not specifically focusing on digital epigraphy, and the I.Sicily – *Inscriptions of Sicily* project (Prag and Chartrand; Chapter 19), which, in addition to a consistent amount of previously undigitized epigraphs, provides original editions based on the principles of reuse, linked data and collaboration, demonstrate the potential of records encoded according to the best practice shared by the scientific community.

The volume is provided with an index, listing terms grouped by: Ancient and Modern Regions and States; Languages and Scripts; Concepts of the epigraphic discipline and related digital practice. Finally, two appendices complement the volume. Appendix A presents an annotated webliography of selected online electronic resources cited in the volume, described according to the Dublin Core Metadata Element Set (Version 1.1). Appendix B is intended for disambiguation and definition of selected concepts from the Index of Concepts, by mapping them to the Library of Congress Subject Headings and the Getty Art and Architecture Thesaurus.

## Reading Path

The deliberate heterogeneity of subjects, focuses and approaches to digital epigraphy represented in this volume, allows a non-sequential fruition of the contributions. However, they are grouped into two main subject areas. These areas, which have been part of the research of DASI itself, enclose, in our opinion, the main issues that digital epigraphy should address in developing a methodology able to provide the validity criteria proper to a discipline.

1. The first part of the volume is focused on data modelling and encoding, which deeply influence the possibility to perform searches on texts including *lacunae* and variants.
  - Various scripts, belonging to different writing systems and often not completely deciphered, pose fundamental issues in relation to data modelling and/or encoding, given the high uncertainty in the attribution of

- phonetic, morphological and semantic values to graphemes and sequences of graphemes.
- Data modelling and encoding are also influenced by the will of creating proper critical editions of epigraphs and the specific functionalities required to meet their criteria.
  - Moreover, different languages, often extinct and not completely understood in their morphology and lexicon, need to be studied from the linguistic point of view, before historical, cultural, sociological and much more interpretation can be derived. Lexica and tools for morphological analysis, specifically developed on the basis of the epigraphic collections digitized, and coping with fragmentarily attested languages, are therefore described.
2. Interoperability and aggregation are fundamental to relate data that would otherwise remain separate, in contrast to the reality they refer to. This second part of the volume is dedicated to the initiatives aimed at fostering aggregation, dissemination and reuse of epigraphic materials. It includes:
- the experiences which point out the need, and tools, for interoperability
  - portals providing “annotated” access to several digitization projects, and proper aggregators
  - and projects which, thanks to interoperability, are clear examples of successful dissemination of inscriptions digitized in different projects.

Although the contributions allow multiple keys to interpretation, and the editors encourage a “personal” fruition, this ordering of the papers aims at suggesting a reading path. This path follows the red thread of the dialectical relationship between the need to represent in the digital environment the features of peculiar epigraphic materials in the most effective way, and the need for strategies to share, disseminate, and make data reusable. In other words, the relationship between the compliance with the theoretic tools and the methodologies developed by each different tradition of studies, and, on the other side, the necessity of adopting a common framework in order to produce commensurable and shareable results in digital epigraphy.

In sum, by crossing a wide, even though not exhaustive, range of experiences, this volume intends to point out the methodological issues which are specific to the application of information technologies to epigraphy. It was not conceived to be a prescriptive work; it does not provide answers, but focuses on problems. Eventually, it aims at stimulating interest and discussion around the challenges that the use of IT has been imposing on epigraphy and on how the digital approach is reshaping the very discipline.

## Acknowledgements

The dedication we have put in the preparation of this volume is our personal homage to the Director of DASI project, Prof. Alessandra Avanzini. Under her guidance, DASI has been for us a time of absorbing study, passionate debate, curious experimentation, continuous rethinking and multidisciplinary encounters with the many people of the DASI research groups of the Scuola Normale Superiore and the University of Pisa, who have shared with us daily enthusiasm and hard work.

We take this opportunity to thank the contributors to this volume, who have dedicated, in spite of their many commitments, time and energies to reflect on common issues and challenges.

Our thanks go also to the editors of De Gruyter, and in particular to Katarzyna Michalak, for their constant assistance.

The FP7 post-grant Open Access Pilot project has provided financial support for the publication of the present volume.

Annamaria De Santis and Irene Rossi



Alessandra Avanzini, Annamaria De Santis and Irene Rossi

# 1 Encoding, Interoperability, Lexicography: Digital Epigraphy Through the Lens of DASI Experience

**Abstract:** This paper describes the main challenges faced and the solutions adopted in the frame of the project DASI – *Digital archive for the study of pre-Islamic Arabian inscriptions*. In particular, it discusses the methodological and technological issues that emerged during the conversion from the CSAI – *Corpus of South Arabian inscriptions* project (a domain-specific, text-based, digital edition conceived at the end of 1990s) to the wider DASI archive for the study of inscriptions in different languages and scripts of ancient Arabia. The paper devotes special attention to: the modelling of data and encoding (XML annotation vs database approach; the conceptual model for the valorisation of the material aspect of the epigraph; the textual encoding for critical editions); interoperability (pros and cons of compliance to standards; harmonization of metadata; openness; semantic interoperability); lexicography (tools for under-resourced languages; translations), with a view to possibly fostering reasoning on best practices in the community of digital epigraphers beyond each specific cultural/linguistic domain.

**Keywords:** data modelling, text encoding, interoperability, lexicography, pre-Islamic Arabia

## 1.1 Digitizing the Epigraphic Heritage of Ancient Arabia: From CSAI to DASI

From the beginning of the first millennium BCE, in the region corresponding roughly to modern Yemen and neighbouring areas in Oman and Saudi Arabia – the so-called *Arabia Felix* of the classical sources – the Ancient South Arabian civilization flourished. During a long history of more than 1,500 years, the Ancient South Arabian four main kingdoms of Maʿīn, Saba, Qataban and Ḥaḍramawt produced a written documentation currently consisting of around 15,000 inscriptions, which constitute the direct textual source for the knowledge of the Ancient South Arabian civilization, as no literary texts have been discovered yet (Avanzini, 2016).

Recognising the need for a systematic collection of this epigraphic heritage, in 1999 Prof. Alessandra Avanzini at the University of Pisa undertook the project of an

---

Alessandra Avanzini, Annamaria De Santis, Università di Pisa  
Irene Rossi, Consiglio Nazionale delle Ricerche, Roma



© 2018 Alessandra Avanzini, Annamaria De Santis and Irene Rossi

This work is licensed under the Creative Commons Attribution 4.0 International license (CC BY 4.0)

online *Corpus of Ancient South Arabian Inscriptions* – CSAI (Avanzini, Lombardini, & Mazzini, 2000). The choice of producing an online curated textual corpus – even before considering its paper edition (Avanzini, 2004) – was determined by several advantages that apply to any cultural domain of study, but that are especially indispensable for those “young” disciplines, whose progress determines a constant re-definition of previous theories. Those advantages are: the updatability and expandability of the collection, the potential improvement of the edition of the sources and of the consultation tools, including full-text retrieval tools, the immediate accessibility of the material – published in scattered, often inaccessible publications, or coming to light from excavations at a fast pace – and its potentially infinite dissemination.

The CSAI archive, realized with the technical support of the Scuola Normale Superiore di Pisa, went online in 2001. Its starting bulk was comprised of some 1,300 texts of the *Corpus of Qatabanic Inscriptions*. The archive content was continuously updated for a decade, so to comprise the whole collection of Qatabanic, Minaic and Ḥaḍramitic inscriptions, plus a number of Sabaic texts – Sabaic being the most consistent South Arabian epigraphic corpus (Figure 1.1).

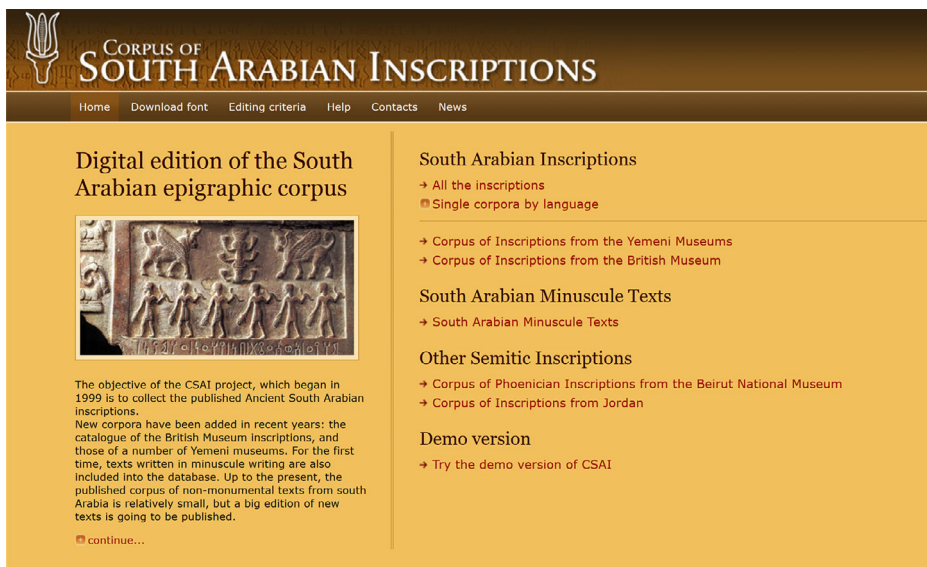


Figure 1.1: CSAI homepage (2010)

Related, funded projects aimed at the cataloguing of not just the Ancient South Arabian, but also the Nabataean and Phoenician collections of inscriptions and

artefacts preserved in museums worldwide,<sup>1</sup> allowed the content of the archive to be enriched. These projects also enhanced a continuous methodological reflection and technical elaboration, allowing a definition of best practice and development of tools for the study of a peculiar documentation, whose state of research is still “fluid”.

It is precisely from this kind of experience, that some ten years later a wider project, the Digital Archive for the study of pre-Islamic Arabian inscriptions (DASI), was conceived and funded with an ERC Advanced Grant awarded to Prof. Avanzini. The objective was to enhance knowledge of the history, language and culture of the whole of ancient Arabia by studying its textual heritage; a heritage that is composed of tens of thousands of inscriptions in the Ancient North Arabian, Ancient South Arabian and Aramaic languages and scripts.

Both the digitization tool and archive’s public website of the CSAI were re-designed, in order to conform to the new research objectives of the DASI project and to the advancements in digital humanities that had occurred during the last decade (cf. in general Schreibman, Siemens, & Unsworth, 2004; Babeu, 2011). The process of re-engineering a system which already contained a large amount of data (around 6,000 inscriptions, with encoded text, metadata, translations, bibliographical references and visual documentation) and the migration of structured data, brought to light a series of methodological and technical issues. Only part of them could be satisfactorily faced.

In the present paper, the main challenges we encountered, the proposed solutions, the still open questions and the prospects we envisage for the future of digital epigraphy – starting from our experience within the DASI project – will be discussed, dealing with three core themes: data modelling and text encoding, harmonization and interoperability, and lexicography.

## 1.2 Data Modelling and Textual Encoding

### 1.2.1 The Data Model: XML vs Database

During the 1990s, textual encoding was successfully experimented with literary sources, and became the standard approach for projects interested in digitizing and annotating texts. The IT system of the CSAI was developed by the “Centro di Ricerche Informatiche per i Beni Culturali” (CRIBeCu) of the Scuola Normale Superiore of Pisa, which had acquired specific know-how in the field of text encoding and had

---

<sup>1</sup> MENCAWAR – Mediterranean Network for Cataloguing and Web Fruition of Ancient Artworks and Inscriptions [<http://arabiantica.humnet.unipi.it/index.php?id=mancawar>]; CASIS – Cataloguing and Fruition of South Arabian Inscriptions through an Informatic Support [<http://arabiantica.humnet.unipi.it/index.php?id=casis>].

developed TRSy (acronym for Text Retrieval System). This was one of the early full-text SGML-XML search engines, able to perform accurate queries on the context (Lini et al., 2004). Metadata and texts of the CSAI inscriptions were recorded in SGML, and later XML files, according to a schema specifically created for CSAI. Indeed, best practice and standards, such as those of TEI and EpiDoC, were not yet widespread, especially in Europe.

This kind of approach, centred on the manipulation of the text, suffered from a range of shortcomings in the description of the text-bearing object and in the management of complementary resources such as bibliographical records and visual documentation. Moreover, the system forced users to handle the XML, often discouraging potential encoders, and did not allow the control of the workflow and the real-time updating of data.

To overcome these limitations, a new system was designed for the DASI project by the staff of the Scuola Normale Superiore, consisting of a web based, relational database enabling a controlled and swift workflow by different levels of authorization for each curatorial role, and uniformity of data by an extensive use of lists of controlled terms, editable by authorized users.

An XML editing module for the textual transcription and encoding of the pre-Islamic Arabian inscriptions was integrated into the database. This is provided with a set of buttons to enter the annotation of all, and only the phenomena considered within the project, ensuring easiness and uniformity of mark-up. The validity and well-formedness of the documents against the schema are granted by preventing elements being entered in incorrect positions, and by managing overlapping of tags through a system of identifiers and couplings between the fragments of the broken elements. The entire content of the database – text encoded and metadata – is then extracted in XML by a web service, in order to construct the dynamic sections of the front-end.<sup>2</sup>

In the context of a “niche” discipline, the design of easy-to-use tools such as the DASI XML editor, as well as the entire data entry system (Figure 1.2), was an effective step towards a wider involvement of scholars in the digitization and curatorial tasks. Moreover, DASI system has proved to be a performing didactic tool in the teaching of epigraphic disciplines and Semitic languages. The virtual keyboard with diacritic characters helps in the transliteration, and the scientific terminology displayed on menus and buttons for textual mark-up suggests coherent definitions to be used to. The process of encoding develops the students’ familiarity with methods and tools of philological and linguistic analysis.<sup>3</sup>

---

<sup>2</sup> [<http://dasi.cnr.it>]. DASI IT system is currently maintained by the CNR Reti e Sistemi Informativi, with the scientific supervision of the CNR Dipartimento Scienze Umane e Sociali, Patrimonio Culturale.

<sup>3</sup> Cf. Bodard & Stoyanova, 2016 for similar experiences in the domain of Classic epigraphy.



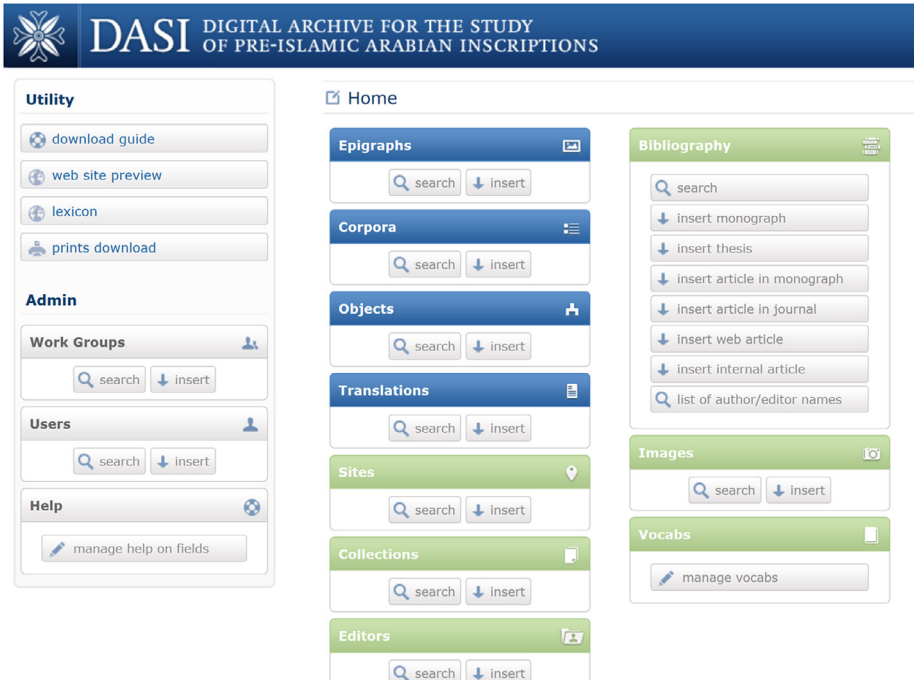


Figure 1.2: DASI data entry interface

### 1.2.2 The Conceptual Model: Text vs Object

Given the obvious focus of the DASI project on the inscriptional text, as any other epigraphic project, the Epigraph entity (see below) is the most articulated one in the conceptual model of the database. Besides the XML editor for textual transcription and annotation, it contains the metadata of the text (on linguistic features, writing, chronology, genre, notes of *apparatus criticus*, general and cultural notes).

Metadata and text of the inscription's translation(s) are recorded in the related Translation entity. Additional entities complete the description with geographic information (Site), visual documentation (Image), references to the history of studies (Bibliography) and indications of curatorial responsibilities within the DASI project (Editor).

The core issue in the conception of the DASI model was the need to account for and valorise the material aspect of the epigraphic document. As stated above, in the traditional encoding approach this proved to be under-represented in comparison to the textual aspect, to such an extent that information on the supports of the inscriptions was encoded as metadata of the text.

Therefore, the innovation in the DASI approach, compared to the CSAI, is the separation of the information concerning the text from that concerning its physical support in two different but related entities. The recording of the archaeological and historic-artistic information on text supports in a dedicated Object entity, allows the additional problem of the multiplication of object records in the case of objects bearing multiple, self-contained texts to be overcome, and the one-to-one relation between the object in the database and the real object to be maintained. Moreover, the autonomy to the Object entity allows to record uninscribed objects, with the additional outcome of enhancing the study of the history of art of pre-Islamic Arabia and valorising specific museums' collections of objects in the DASI archive.

The DASI website reflects the text-object distinction, via the two main indexes of *corpora* and *collections* that group texts and supports on the basis of their linguistic attribution or current deposit respectively. This has proved extremely important for the preservation and valorisation of the Yemeni cultural heritage, as some of the museums' collections catalogued in DASI have undergone serious damage or pillage, or were entirely destroyed during the ongoing war.<sup>4</sup> Securing the existence of digital copies of objects at risk from environmental or human factors is today of primary importance. We believe that their description as well as their visual documentation – and the open access and re-usability of both – should be among the major concerns of projects involved in the digitization of cultural heritage, for preservation purposes.

The distinction proposed in the DASI model between texts and supports, though suitable from the conceptual and practical points of view, has its limits due to the strict relation between them (e.g. the spatial relations among components of the text, the distribution of text on the support, the relation of the texts with the iconographic elements and decoration), and with the communicative context. The case of the monograms is emblematic. The monogram is not an abbreviation inside the text, but a combination of signs decorating an object (Figure 1.3). The same monogram may occur engraved next to a text or even without a linear inscription. In many cases, the name the monogram refers to is unknown, because some letters can be omitted or incorporated into the shape of other letters, there being no way to reconstruct their correct order in these symbolic representations. Therefore, are monograms inscriptions, or rather decorations, of objects? Should they be encoded in the Epigraph or described in the Object?

A further example of the relation between texts and supports is the mention within the epigraphic text of the type of object on which it is inscribed. The correspondence between the term and its material signifier is extremely relevant for the improvement of the knowledge of both the pre-Islamic Arabian languages and the material culture. However, the data model adopted does not allow for a direct correlation between

---

<sup>4</sup> [<http://en.unesco.org/galleries/heritage-risk-yemen>].

them, nor between parts of the text and their visual documentation, which may be improved through the tagging of images.



**Figure 1.3:** Early Sabaic boustrophedon inscription with monogram and symbols (MŞM 149)

### 1.2.3 Encoding for Curated Digital Editions: In-Line vs External Apparatus Criticus

The XML editor integrated in the DASI data entry system (Figure 1.4) allows encoding of texts in compliance with the EpiDoc subset of the TEI<sup>5</sup> standard (Elliott et al., 2007–2016). The annotated phenomena are linguistic (onomastic, grammatical), philological (lacunae, restorations, corrections, etc.), descriptive of the relation between text and support (line breaks, text turning around the object) or of the internal structure of the text (genealogies, eponyms), etc. The critical notes are collected in a separate

<sup>5</sup> Text Encoding Initiative [<http://www.tei-c.org/>].

section and refer to the concerned text by the indication of the corresponding line of the transliteration – a traditional approach of managing the *apparatus criticus* that has been inherited from the project CSAI.

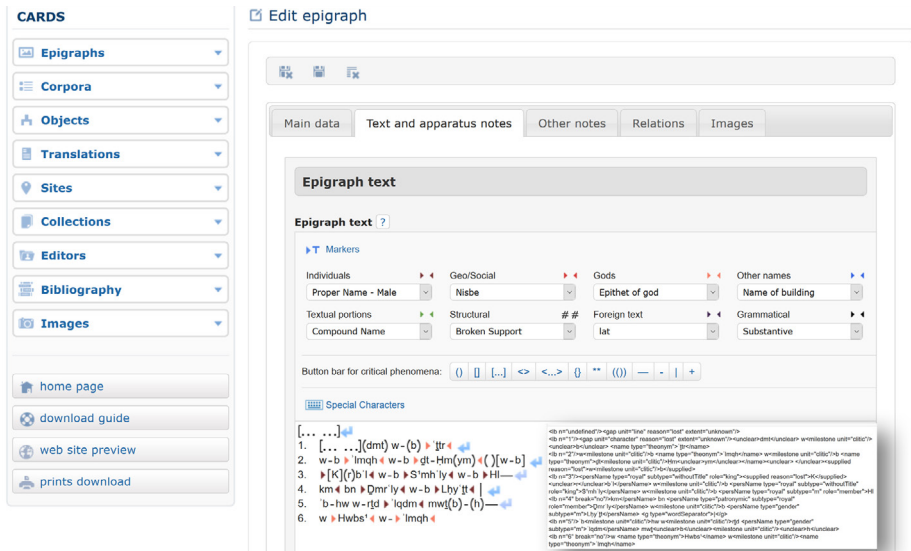


Figure 1.4: DASI XML editor

The solution many projects have adopted in order to valorize the apparatus notes is the encoding of the text contained in them, and its referencing to the corresponding section of the epigraph transcription. The alternative solution is the insertion of the *apparatus criticus* in-line, directly in the transcription's annotation. This is particularly interesting, as it allows retrieval, through a textual search, of all the possible readings/interpretations of a textual passage, or the renderings of the texts suggested by different editors. Indeed, the *<app>* with *<lem>* and *<rdg>* elements have been used in the DASI XML editor to encode variants of uncertain readings or of restorations, or of linguistic (mainly onomastic) interpretations, when none of them could be discarded.<sup>6</sup> As it is apparent, the main concern in the DASI in-line encoding of variants is not so much to retrieve single variants of words, as to retrieve them

<sup>6</sup> These elements were created in the TEI to encode the variants occurring in a work's multiple witnesses, as in the case of manuscripts. However, their semantic value can be applied to encode information on different critical editions of one epigraph, because the strong emphasis on the physical nature of an epigraph leads to consider each inscription as a unique specimen. This solution is presently suggested also in [http://www.stoa.org/epidoc/gl/latest/supp-app-inline.html]. EpiDoc guidelines in general are available at [http://www.stoa.org/epidoc/gl/latest/].

within a specific context, consisting in the portion of text preceding and following the text characterized by variants.

The tool for combined searches on text and extra-textual data provided in the DASI website allows queries on words or word patterns within a phrase, with the possibility of setting the maximum number of words that can intervene between the first and the last words searched for (Avanzini, Priolella, & Rossi, 2014). The search can be restricted to lexical or onomastic results – even within a particular onomastic category. The adoption of an in-line approach in recording the *apparatus criticus* of the inscriptions would exponentially augment the potential of such a search tool.

However, to encode the *apparatus criticus* in-line at such a level of detail as to provide an “encoded history of study” of an inscription (i.e. rendering on one single file the interventions applied by different scholars in their own edition of the text) is a very long and complex task. Moreover, it entails the risk of over-tagging the transliteration of the text by applying too many “layers” on it. On the other hand, providing several files for the different editions of the same inscription, to be then grouped within aggregators, is a viable solution, but it limits the potentialities offered by a digital edition.

## 1.3 Interoperability

### 1.3.1 Text Encoding and Representation: Standards vs Specificities

All of the scripts used to write down the inscriptions considered within the DASI project (Ancient South Arabian, Ancient North Arabian and Aramaic varieties) are alphabetic. In Southern Arabia a geometric, monumental writing is evidenced since the 9<sup>th</sup>–8<sup>th</sup> century BCE by the “public” inscriptions: each letter is graphically separated from the adjacent ones and the division between the orthographic units (which, as typical of the Semitic languages, can be composed by a main word plus affixes for clitic pronouns, conjunctions, particles) is marked by a vertical trait. A “cursive” writing was also in use to record private, movable or archival texts on wooden sticks (contracts, lists of goods, correspondence, school exercises, etc.). As the majority of Semitic scripts, the ductus of writing is normally right-to-left, although in ancient South Arabia there are a considerable number of boustrophedon inscriptions as well. The Ancient North Arabian texts – except for a few hundred “monumental” texts from major settlements – consist mainly of graffiti left by nomadic people on desert rocks, and their direction of script is much more varied, sometimes even circular.

The inscriptional text is entered in the DASI XML module in Latin transliteration, using the UTF-8 set of the Unicode standard. The transliteration of Semitic phonemes in Latin characters implies the addition of diacritical marks (like underdots) to the letters and therefore discourages the representation of editorial phenomena according to the Leiden conventions: the latter, elaborated in the frame of Classical philology

and recommended by EpiDoc, visualise the uncertain reading of signs precisely by dots under the letters.

More generally, the DASI project has adhered to the TEI-EpiDoc standard to encode texts, with some limitations imposed by: the need to comply with the specific tradition of studies (choice of phenomena to annotate), the inheritance of the CSAI custom-made encoding schema (already applied to some 6,000 inscriptions) and, related to this, the peculiar interests of the project (like the linguistic, more than prosopographical focus on onomastics). This process of mark-up conversion and the effort towards the alignment to a standard have shown their potentialities in terms of content rethinking and redefinition, and at the same time the need to safeguard as much as possible the specificities proper to each cultural domain and tradition of studies, in order not to lose peculiarities, profoundness and nuances (Avanzini et al., 2016).

### 1.3.2 Harmonization of Metadata

As explained above, the DASI encoding of texts does not fully comply to the EpiDoc standard's recommendations as regards some transcription phenomena and editorial interventions, and for the encoding of onomastics. However, particular attention has been paid to the harmonization of those metadata elements that entail a reference to structured terminologies. Indeed, the tradition and the state of the art in a discipline exert their influence above all in the classifications that stand at the basis of the knowledge organization systems.

This is exemplified by the lists of controlled terms related to the textual typology and the type of object, which best show the progress in the understanding of the peculiarities of the pre-Islamic Arabian textual tradition and material culture (Avanzini, Prioletta, & Rossi, 2014). The three main typologies of inscriptions – i.e. dedications to the gods, celebrations of construction activities, and legal/administrative regulations – are distinguished by specific formulary patterns (lexical items – in particular the main verb of the inscription, which is the fulcrum of the action described throughout the text – and syntactic features) and very rigid textual structures (the order of the text sections). These were replicated through the centuries, with few areal and chronological variants, and rarely conceded space to the insertion of digressions or to the combination of different textual typologies in the same inscription. The texts encoded in the DASI archive are classified on the basis of those fixed textual models. The comparison with terminologies used in other projects, such as those harmonized in the project EAGLE, has pointed out that some of the entries have exact matches, others are just related to some terms, and the remaining ones have no match at all. This is because different criteria have guided the creation of such classifications and therefore of the vocabularies in use.

Even internally, the DASI project has faced the issue of managing a diversified documentation.<sup>7</sup> The textual encoding accounts for all of the three main language corpora considered within the project (Ancient South Arabian, Ancient North Arabian, and Aramaic), though with obvious compromises as regards specific grammatical features and definitions for each language. It was more difficult to find shared solutions for metadata. For instance, the CSAI project had catalogued and annotated information that was especially relevant to the comprehension of the “monumental” inscriptions (the majority of Ancient South Arabian texts), while most of the Ancient North Arabian inscriptions consist of graffiti. The two categories of texts considerably distance themselves with respect to their scope, audience, authorship, context, etc.; therefore the information that one wants to point out and extract to enhance their study is different. For instance, much attention has to be paid to the artistic description of the support of a monumental inscription, whilst the technique of incision and the relative disposition of texts on a rock are essential information to describe graffiti.

As regards the physical supports of the texts, the specimen that DASI has collected demonstrates its own peculiarities. For instance, stelae make a large part of the artefacts catalogued. Common terminologies, such as the Getty Art & Architecture Thesaurus,<sup>8</sup> include only one term to classify them, but the South Arabian stelae have different, codified morphological and iconographic characteristics that are fundamental (as much as their texts) for the identification of their area of production, dating and function.<sup>9</sup>

Even for those entries that have exact matches, further subcategories may be required to provide specifications useful to scholars interested in a particular domain. In South Arabia, for instance, bases can be found as support to statues, sculptures of heads and stelae. Their morphological and functional – i.e. communicative, not only material – features, as well as the geographical and chronological distribution, may vary considerably. Is it possible to increase the granularity of the shared terminologies without reproducing the domain-specific typologies of the classes of materials? For instance, let us consider the bases of Ancient South Arabian statuettes, which have been found in temples for propitiatory and votive aims. We would consider it inappropriate to map the South Arabian bases to a concept having such a domain-

---

7 When designing the metadata and the tags of the XML editor, the project benefited from the collaboration of colleagues at the CNRS-UMR Orient & Méditerranée as regards the Aramaic corpus, and at the University of Oxford as regards the Ancient North Arabian corpus (see Chapter 8 in this volume).

8 [<http://www.getty.edu/research/tools/vocabularies/aat/>].

9 For instance, large, rectangular stelae with a decoration of ibexes and *bucrania* framing the text, always bear dedicatory texts and are typical of the Sabaic and Minaic areas, especially in ancient times. Small trapezoidal aniconic stelae whose base is inserted on an inscribed plinth, as well as rectangular, beautifully carved stelae with the representation of the deceased's bust and his/her name inscribed below the figure, usually bear Qatabanian funerary texts.

driven definition as the bases of statues in the Classic world, which are placed in the public, civic space with honorary function.

### 1.3.3 Openness and Semantic Interoperability

In relation to the public funding of the project and the policy adopted by the EC on Open Access to publications and research data, the DASI project has made available the entire archive in open-access modality. The DASI repository allows service providers to harvest its records through the OAI-PMH protocol (Avanzini et al., 2015).<sup>10</sup> As the archive is not an aggregator in the strict sense, the DASI project has developed a general data model able to convey an accurate description of the material support, the historical and geographic context, and the textual content of the pre-Islamic inscriptions of the Arabian Peninsula, but not a proper schema. Therefore, the key point has been mapping the DASI data model to the DC elements set, as required by the OAI-PMH protocol, and to the EDM in order to expose records to the Europeana aggregation service, in addition to the mentioned EpiDoc subset.

A further step to achieve semantic interoperability,<sup>11</sup> in addition to interoperability at the repository level and at the record level, is related to the names of individuals and places. The DASI encoding of onomastic phenomena is detailed and articulate. However, for the time being, it has had a linguistic objective rather than a prosopographical one. The royal onomastics is easily recognizable and extremely repetitive, as it was probably taken on with the investiture. Genealogies of kings are therefore rather evanescent, so much to suggest that the institution represented was more important than the individual king, at least until the last centuries BCE (Avanzini, 2016, pp. 53–57). Then, it is difficult and highly hypothetical to identify a single person, place him/her over time, and relate with certain attestations. Nevertheless, it would be worth seeking to do this for the main historical figures and for some periods, for instance when inscriptions begin to be dated and therefore the identification of individuals is less tentative.

Similar considerations could be made about place names. The DASI onomastic lists include about 3,600 names of geographical, social and political entities that have been tagged in the epigraphs: elements of the natural and the human landscape, entire settlements and single artifacts (buildings and monuments), political and social entities (states, tribes, families) which have relations with the territory. Furthermore,

---

<sup>10</sup> DASI repository [<http://dasi.cnr.it/de/cgi-bin/dasi-oai-x.pl?verb=Identify>].

<sup>11</sup> DASI does not apply a frankly semantic approach from the technical point of view, even though the distinction between the physical carrier and the text inscribed in the data model is an implicit result of that way of conceptualizing. However, the mapping of its data to the Europeana Data Model goes in that direction.



archaeological data related to nearly 400 sites, origin or provenance of inscriptions, have been collected: modern and ancient toponyms, including Classical names; country, geographical area and present governorate, coordinates and related accuracy; types of the findings, architectural structures and monuments; chronology; description, history of research; bibliography. Each “Site” record may be linked to the other ones, thus representing the spatial relations among them. A gazetteer is in preparation, which will allow identification and description of all the above-mentioned geographical entities and represent the semantic relations (hierarchy, equivalence and association) among them, in addition to the spatial ones, directly inferred by the primary (epigraphs) or secondary sources (bibliography). This is of particular importance when their actual locations or identification are still unknown.

The difficulties in the historical reconstruction of the pre-Islamic Arabian civilizations are especially apparent at a chronological level, so that the DASI inscriptions are dated to wide periods of three/four centuries. However, as the historical understanding moves forward – and at least for the dated inscriptions since the end of the 1<sup>st</sup> millennium BCE – an attempt at the semantic interoperability at a chronological level has been envisaged, in connection with the PeriodO project (see Rabinowitz, Shaw, & Golden in this volume).

## 1.4 Lexicography

### 1.4.1 Approach to Under-Resourced Languages

Interoperability at a linguistic level across different corpora is a desideratum. The goal of providing useful tools for the research on each of the main corpora that make up the DASI archive (Ancient South Arabian, Ancient North Arabian and Aramaic) has been reached. However, a major issue is still to be approached: whether, to what extent and how to enable combined queries on textual content across documentation in different linguistic families. In fact, not only do these corpora have their own peculiarities (e.g. in terms of language, script, or periodization) that would entail partial or potentially false results, but they also have specific traditions of studies strongly conditioning approaches, methods and definitions. The mapping of grammatical (to a lesser extent) and mainly semantic features of different languages could be one of the ways, though not straightforward and immediate, to facilitate cross-searches on them.<sup>12</sup>

---

<sup>12</sup> During the revision process of the present volume, two books on similar topics have been published. For recent developments and updated bibliography, refer to Juloux, Gansell, & Di Ludovico, 2018, for semantic approach to digital epigraphy of the ancient Near East, and to Cotticelli-Kurras & Giusfredi, 2018, for relations between computational linguistics and digital philology.

In this sense, the DASI project is implementing lexica of the languages attested in pre-Islamic Southern Arabia, whose current state of knowledge is still very fluid. The development of lexicographic tools starting from the dataset of a digital archive is the ideal situation to advance the research on such fragmentarily attested languages, whose dictionaries and grammatical studies need constant updating due to the growth of sources that are brought to the attention of scholars.

The DASI Lexicon tool has its starting point in the list of words (excluding onomastic items) extracted from the texts encoded (Avanzini et al., 2015). Each of the word forms, corresponding to the items of the words' lists, is retrieved within the contexts of occurrence in the single inscription. One or more rows of translation are in turn linked to each occurrence. Word forms can be assigned, individually or in groups, to a root. While assigning a root, users attribute morphological, part-of-speech (PoS) analysis and translation to the word form. Each word is thus defined, and potential homographic forms are disambiguated (Figure 1.5).

The screenshot displays the DASI Lexicon web interface. At the top, there is a navigation bar with 'Admin', 'Edit', 'Preview', and 'Recycler' buttons. Below this is a search and filter section with buttons for 'link status', 'grammar', 'word form contains', and 'root contains'. A table lists word forms and their occurrences. A modal window titled 'Lexicon - Occurrence editing' is open, showing details for the root 'ghbn' and its associated word forms.

word form	occurrences
ghb	<input type="checkbox"/> DHB (watered valley) / Substantive - Construct - Masculine singular 2 occ. <input type="checkbox"/> DHB (bronze) / Substantive - Unclear - Masculine singular 1 occ.
ghbn	<input type="checkbox"/> DHB (bronze) / Substantive - Emphatic/defined by article - Masculine singular 3 occ. <input type="checkbox"/> DHB (watered valley) / Substantive - Emphatic/defined by article - Masculine singular 1 occ.

The modal window 'Lexicon - Occurrence editing' shows the root 'ghbn' and its associated word forms: 'DHB (watered valley) / Substantive - Construct - Masculine singular' and 'DHB (bronze) / Substantive - Unclear - Masculine singular'. It also includes a section for 'Insert meanings for a new root' with a text input field and a search icon.

Figure 1.5: DASI Lexicon

This clearly manual approach is mainly related to the constraints of the languages concerned. In fact, the pre-Islamic Arabian languages share with the other, also current, Semitic languages a morphological ambiguity, that is itself a challenge to computational approaches (see Multhoff in this volume). In addition to the typical case of more than one analysis for a given word form, there can be different graphical renderings (spellings) for the same word. Moreover, the high number of *hapax legomena* causes data sparsity. Finally, the small scale of the annotated corpora discourages from effectively driving automatic lexical acquisition.

On the other hand, the remarkable repetitiveness of the texts in relation to the formulaic contexts, suggested not to encode the grammatical and lexicographic information directly on the texts, in order to avoid repeating the editing of all the occurrences. By assigning at one time the same semantic and grammatical analysis to multiple occurrences of a lexical item at a further level, the Lexicon allows completion of the lexicographic work and its potential revision, following the advancements in research, in a reasonable lapse of time.

#### 1.4.2 Translations

While a growing number of digital archives is developing the lexicographic aspect, the majority of them do not include translations, being conceived as traditional collections of primary sources in electronic form.<sup>13</sup> The project DASI, aiming at the study and the dissemination of the ancient culture of Arabia through the analysis of its epigraphic heritage, conceived its archive as a digital mean for publishing and browsing critical editions of texts, provided with translations, for a better and wider appreciation of their content. More than one translation, often in different languages depending on its bibliographic source, may be linked to one epigraph.

This wealth of data, which takes into account the interpretations by different scholars, however, is not yet encoded. The correspondence between a line of the text and a line of the related translation is a best practice followed by the editor of each item, even though a comprehensible rendering of the concept expressed in the original text often prevents this correspondence, for syntactic reasons. Notwithstanding an effort towards homogenization among the contributions by different editors (at least those directly involved into the project), a strict relation between a word and its translation is also undermined by the semantic differences of words deriving from the same root, or by the semantic nuances one word can acquire on the basis of its context, or more generally by the different morpho-syntactic rules of each language.

Having said that, translations are a key point for searches on texts in different languages. Given the multilingualism of the DASI inscriptional records, translations are going to deserve further methodological reflection and technological effort.

---

<sup>13</sup> The awareness of the importance of translations is emerging and several projects, such as AIO “Attic Inscriptions Online” and EAGLE (see Chapter 17 in this volume), have spent much effort on translations to allow a larger public to access epigraphic sources in extinct languages.

## 1.5 Conclusions and General Remarks

The issues discussed in this paper highlight the efforts that long-lasting digital projects have to make in order to be coherent with their very digital nature, which grants, and at the same requires, constant updating and improving of data, tools and practices.

What epigraphic disciplines, and in general the humanities, are experiencing nowadays, is the contraction both of people engaged in those studies and of funding of projects, creating a vicious circle. Particularly for projects that obtained conspicuous funds, allowing the creation of a large team and undertaking a wide range of initiatives, the abrupt end of short-term grants implies a stalemate. Given the additional efforts required by such a scientific production with respect to more “traditional” outcomes, a major appraisal of the digital products, and specifically of curated editions, in the evaluation process that research and academic staff are subjected to, would stimulate a wider engagement of scholars and early-stage researchers, ensuring sustainability of digital humanities initiatives.

Notwithstanding those apparent difficulties, “young”, “niche” domains of studies, such as the one described in the paper, especially need digital tools, as their state of research entails a continuous production of fresh knowledge and review of theories, and at the same time they are likely to boost the discussion on the perspectives of the digital approach to scientific disciplines – epigraphy in our case – by bringing unexpected issues to the forefront.

**Acknowledgements:** The research leading to the results presented in this paper has received funding from the European Research Council under the European Union’s Seventh Framework Programme (FP7/2007-2013)/ERC grant agreement n° 269774.

## Bibliography

- Avanzini, A. (2004). *Corpus of South Arabian Inscriptions. I-III: Qatabanic, Marginal Qatabanic, Awsanite Inscriptions* (Arabia Antica 2). Pisa: Edizioni Plus-Università di Pisa.
- Avanzini, A. (2016). *By Land and by Sea: A History of South Arabia before Islam recounted from Inscriptions* (Arabia Antica 10). Roma: L’«Erma» di Bretschneider.
- Avanzini, A., De Santis, A., Gallo, M., Marotta, D., & Rossi, I. (2015). Computational Lexicography and Digital Epigraphy: Building digital lexica of fragmentary attested languages in the Project DASI. In G. Guidi, R. Scopigno, J.C. Torres, & H. Graf (Eds.), *2015 Digital Heritage International Congress* (pp. 405–408). New York: IEEE. doi: 10.1109/DigitalHeritage.2015.7419535
- Avanzini, A., De Santis, A., Marotta, D., & Rossi, I. (2014). Between harmonization and peculiarities of scientific domains: Digitizing the epigraphic heritage of pre-Islamic Arabia in the project DASI. In S. Orlandi, R. Santucci, V. Casarosa, & P.M. Liuzzo (Eds.), *Information Technologies for Epigraphy and Cultural Heritage: Proceedings of the First EAGLE International Conference* (Serie antichistica. Collana Convegni 26) (pp. 69–93). Roma: Sapienza Università Editrice. Retrieved from [<https://www.eagle-network.eu/wp-content/uploads/2015/01/Paris-Conference-Proceedings.pdf>], 2017/11/30.

- Avanzini, A., De Santis, A., Marotta, D., & Rossi, I. (2016). Is still Arabia at the margins of digital epigraphy? Challenges in the digitization of the pre-Islamic inscriptions in the project DASI. In A.E. Felle & A. Rocco (Eds.), *Off the beaten Track: Epigraphy at the Borders. Proceedings of the VI EAGLE International Meeting (24-25 September 2015, Bari, Italy)* (pp. 46–59). Oxford: Archaeopress. Retrieved from [http://www.archaeopress.com/ArchaeopressShop/Public/download.asp?id={E7B2AAC6-9986-4C41-9842-6AA93BE7ACD9}], 2017/11/30.
- Avanzini, A., Lombardini, D., & Mazzini, G. (2000). Corpus of South Arabian Inscriptions. La pubblicazione integrale del corpus sudarabico qatabanico. *Bollettino d'Informazioni del Centro Informatico per i Beni Culturali della Scuola Normale Superiore*, 10.
- Avanzini, A., Prioleta, A., & Rossi, I. (2014). The Digital Archive for the Study of Pre-Islamic Arabian Inscriptions: An ERC project. *Proceedings of the Seminar for Arabian Studies*, 44, 15–24.
- Babeu, A. (2011). “Rome Wasn’t Digitized in a Day”: Building a Cyberinfrastructure for Digital Classicists. Washington: Council on Library and Information Resources. Retrieved from [https://www.clir.org/wp-content/uploads/sites/6/pub150.pdf], 2017/11/30.
- Bodard, G. & Stoyanova, S. (2016). Epigraphers and Encoders: Strategies for Teaching and Learning Digital Epigraphy. In G. Bodard & M. Romanello (Eds.), *Digital Classics Outside the Echo-Chamber: Teaching, Knowledge Exchange & Public Engagement* (pp. 51–68). London: Ubiquity Press. doi: 10.5334/bat.d
- Cotticelli-Kurras, P. & Giusfredi, F. (Eds.). (2018). *Formal Representation and the Digital Humanities*. Cambridge: Cambridge Scholars Publishing.
- Elliott, T., Bodard, G., Mylonas, E., Stoyanova, S., Tupman, C., Vanderbilt, S. et al. (2007-2016). *EpiDoc Guidelines: Ancient documents in TEI XML (Version 8)*. Retrieved from [http://www.stoa.org/epidoc/gl/latest/], 2017/11/30.
- Bigot Juloux, V., Gansell, A.R., & Di Ludovico, A. (Eds.). (2018). *CyberResearch on the Ancient Near East and Neighboring Regions. Case Studies on Archaeological Data, Objects, Texts, and Digital Archiving*. Leiden: Brill.
- Lini, L., Lombardini, D., Paoli, M., Colazzo, D., & Sartiani, C. (2004). XTReSy: A Text Retrieval System for XML documents. In D. Buzzetti, G. Pancaldi, & H. Short (Eds.), *Augmenting Comprehension: Digital Tools for the History of Ideas*. London: Office for Humanities Communication Publications, King’s College.
- Schreibman, S., Siemens, R., & Unsworth, J. (Eds.). (2004). *A Companion to Digital Humanities*. Oxford: Blackwell. Retrieved from [http://www.digitalhumanities.org/companion/], 2017/11/30.
- TEI Consortium (Eds.). *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. TEI Consortium. Retrieved from [http://www.tei-c.org/Guidelines/P5/], 2017/11/30.





## Part I: Data Modelling and Encoding for Curated Editions and Linguistic Study





Christiane Zimmermann, Kerstin Kazzazi and Jens-Uwe Bahr

## 2 Methodological, Structural and Technical Challenges of a German-English Runic/RuneS Database<sup>1</sup>

**Abstract:** The joint academy project *RuneS* is designing an image-based, multi-relational, bilingual database of all known runic inscriptions. Every runic *find* is characterized according to criteria such as *find-spot*, *find year*, *material*, etc., and transliteration and interpretation of the inscription are given (*RuneS 1.0*). Each individual runic *graph* is then documented by a snippet and classified according to a tailor-made typology as a *graph type (variant)* at a formal level; these are subsequently analysed graphemically at the functional level (*RuneS 2.0*). In a second section, data on the runic texts and their socio-historical function are added (*RuneS 3.0*). The database has the two-fold function of serving as a project-internal research tool as well as constituting a searchable digital corpus available to the wider academic public, for instance for comparative studies involving other epigraphic traditions. The structural, technical and terminological challenges posed by this design are highlighted with screenshots of the first module of the database, the *find* fields.

**Keywords:** runic graph, graph type (variant), grapheme analysis, multi-relational bilingual database, image-based visual documentation/classification

### 2.1 Introduction

#### 2.1.1 The Main Research Areas and the Specific Profile of *RuneS*

The acronym *RuneS* is a discontinuous shortening from the German syntagm “Runische Schriftlichkeit”, i.e. Eng. “Runic writing”, the first two words of the full project title, namely: “Runic writing in the Germanic languages”. The project is funded by the Union of the German Academies of Sciences and was accepted as a

---

<sup>1</sup> This article is based on two conference papers: Christiane Zimmermann and Kerstin Kazzazi: “Structural and Terminological Challenges of a German-English *RuneS*-Database”, presented at a conference in Nyköping in 2014, and Jens-Uwe Bahr: “Digitale Erfassung und Beschreibung runischer Funde und Schriftzeichen im *RuneS*-Projekt”, presented in Mainz in 2017.

---

Christiane Zimmermann, Kerstin Kazzazi, Jens-Uwe Bahr, Akademie der Wissenschaften zu Göttingen



long-term research undertaking by the German Academy of Sciences and Humanities in Göttingen in 2010.<sup>2</sup>

There are two principal domains of investigation, carried out during a research period of six years each: the first focuses on *Runic Graphemics*; the second research topic is *Runic Text Grammar and Pragmatics*. Thus, the project brings together two important current research perspectives on script and writing. In so doing it also draws on results of earlier orality-literacy research carried out by Peter Koch and Wulf Oesterreicher (1985) who introduced this two-fold approach to script and writing analysis.<sup>3</sup>

The investigations are guided by the concept of the runic script as a *system* or, to be more precise, as a *group of writing systems* that evolved in various ways over the centuries, fulfilling different communicative functions within the respective historical societies. The theoretical approach may therefore be characterized as “systematic”, “comprehensive”, and “context-sensitive”. It is the aim of the project to transcend the boundaries of the traditionally separate research perspectives focusing on the groups of runic writing (i.e. the inscriptions using the so-called rune rows of the older *futhork*, the Anglo-Frisian *futhorc*, the Viking Age *futhork/futhork*, and the medieval Scandinavian runic systems). This aim will be achieved by subjecting the respective inscriptions to uniform methods of investigation, thus making them comparable and productive as sources for comparative studies into the “how” and “why” of specific developments and changes of the runic writing systems in use.

### 2.1.2 *RuneS* and Digital Epigraphy

One of the aims of the project is the creation of a joint database containing data on all runic inscriptions, i.e. runic epigraphic material, as well as on the so-called *Runica Manuscripta*, i.e. non-epigraphic use of runes in manuscripts.

The backbone of this database is formed by basic data on the runic finds; these include information on the different types of runic objects, on the find-spots and contexts, and on the dating of the objects and the inscriptions they bear. Two extensions comprising graphemic information, on the one hand (i.e. description and classification of runic graphs and further details on their linguistic function and systemic character/affiliation), and text grammatical/pragmatic data, on the other, are connected to this backbone.

---

<sup>2</sup> For an overview of the overall research plan and the project structure cf. our homepage [<http://runes.adw-goe.de>].

<sup>3</sup> For a more detailed outline cf. the presentation held at the International Conference on Runes and Runic Inscriptions in Oslo 2010 [<https://www.khm.uio.no/english/research/publications/7th-symposium-preprints/runic-writing-scan.pdf>].

This database is designed both as a source of information for the academic community and as a working tool for our own research within the project. It is therefore an ongoing process, with all new research results being entered into the database continuously.

The first module of this database, termed *find* or *RuneS 1.0* (see below), is nearing completion, containing approximately 8,000 entries. The module on *runic graphemics* is currently in the development stage.

The challenges in creating this digital epigraphy database included both structural and conceptual issues, some of which will be addressed in the following sections of this contribution.

### 2.1.3 Why is a Digital *RuneS* Database Necessary?

The use of a digital database has a number of advantages over the classical, analog approach, which are of vital importance for the aims of the project and for our joint venture setting, involving several geographically distant research units. On the one hand, the rather rigid structure of a database ensures the consistency of the analytical approach over a longer period of research. On the other, it allows for working with a large amount of complex, and in various ways interlinked, data of different types such as photographs, snippets, representations and descriptions of individual graphs as well as their abstractions in form of graphemes. The variability in combining these miscellaneous pieces of data (e.g., the runic graph-type variants, the material of the runic object, its socio-cultural function, its dating and provenance resp. find-spot) allows for formulating and verifying/falsifying different hypotheses, e.g., on the socio-cultural distribution and development of runic letter forms and on their systematic interplay. All these aspects and functions are prerequisites and necessary to ensure the “systematic”, “comprehensive”, and “context-sensitive” theoretical approach of the project. Our objective is to address, in this way, questions regarding, e.g.: the connection of the use of rounded vs. angular rune forms (along with other aspects of the graph form) with time, place and/or material; the emergence and development of word, syntagm and sentence separation in runic writing; the continuity and change of functions of runic writing in the historical societies; or questions such as the “de-reification” of the inscribed object in Scandinavia, where the inscriptions increasingly occur on purely functional carriers such as rune sticks.

Additionally, the implementation of a database, not only as a documentary device for storing and displaying the results of research, but also as a fundamental working tool for the investigations themselves, makes it necessary to reflect, in a more consistent and systematic way, the overall structure and the individual steps of the envisaged research plan. This research plan has served as the blueprint for the basic database design and development, itself, however, being modified and adjusted in the process. Questions of systematization of the required research data and of

terminological standardization had to be addressed. In the following sections, we will present several of these steps and some related issues, also regarding bilingual terminology.

## 2.2 Design of the Database

### 2.2.1 Design of the Database – Step Zero: Basic Considerations

As a prerequisite for the basic design of a database we felt that three factors should be taken into account: 1. the kind and relation of the data to be collated, 2. the required data relations and database queries, 3. the issue of the flexibility to respond to future research questions and different user groups.

In addition to its function as a fundamental working tool for the research teams, it was clear from the outset that the *RuneS* database should be made available online to a broader (academic) public. Thus, the structure and format of the data entries needed to allow for bilingual access in both German and English.

Therefore, the database structure for the research module on *runic graphemics* (*RuneS* 2.0) has to be designed along the lines of the following questions:

1. Which kinds of data are relevant for the graphemic analyses?
2. Which database structure is required to allow for the necessary combinations and different searches of the data to answer the relevant research questions?
3. How does the necessity to provide bilingual German-English data sets and interfaces influence the design of the database?

### 2.2.2 Design of the Database – Step One: Type of Data?

The main goal of the research module on *runic graphemics* is to document, describe and explain the process of runic writing, specific phenomena of runic writing systems and the diachronic and diatopic development of the runic script. On the one hand, this requires a systematic formal description and, on the other, a functional analysis of the signs recorded on the runic monuments of the sub-corpora involved. This means:

- Graphs and graphic variations should first be described and classified regarding their shape only, without reference to their function.
- The graphic variants should then be subjected to a functional analysis.
- In the course of the functional analysis, graphic variation should be studied with regard to various relations and dependencies: one is the relation to the phonemic system(s) of the language(s) under consideration. Phonemically non-distinctive variation should then be submitted to context-sensitive analyses to discover further distribution patterns based on context factors. A case in point would be the distribution of the Old English *s*-allographs: whereas the *s*-allographs of the

5<sup>th</sup> and 6<sup>th</sup> centuries belong to the so-called *diagonal type* (cf. Waxenberger, 2000) with one stave forming a zigzag line (tri- or tetra-partite), it is only in the 7<sup>th</sup>–9<sup>th</sup> centuries that the so called *bookhand-s* (allograph) was used (Page, 1973, p. 50, fn. 6).

The clear distinction between a purely formal description and a functional analysis of the graphs may at first glance seem somewhat overly detailed; after all, the function – i.e. the sound value of most runes – has, for most of the graphs, been determined long ago, at least at a phonemic level. However, a digital database provides an opportunity to go beyond common runological knowledge in several ways:

- Providing a comprehensive and uniform basis for the investigations within the project itself: it is the aim of the project to transcend the boundaries of the traditionally separate research perspectives focusing on the different groups of runic writing (see above, 2.1.1), to make them productive as sources for internal comparative studies. This requires a uniform and consistent description language, transcending the boundaries of the various description and classification systems currently in use for the different runic sub-corpora.
- Providing a comprehensive and systematic basis for further and new approaches within runology: in order to be able to set specific graphic variants in relation to different types of potentially influencing contextual factors such as time, place, material etc., it is necessary to document the relevant context data and thus give as full a description as possible of the runic monuments under consideration. This also includes the purely formal make-up of the individual graphs.
- Providing a comprehensive and systematic, strictly formal description for comparative epigraphic studies: for scholars working with scripts other than runes, the starting point for comparison would be the overall shape of the elements the runic symbols are composed of. By providing as the starting point the formal description, it is hoped the digital database will develop into a vehicle for our overall aim of contextualizing runic research in the wider field of epigraphic and general writing research.
- Providing a solid basis for the description and classification of new runic symbols: since the *RuneS* project started in 2010, two new, i.e. hitherto completely unknown Old English runes, have appeared in two new runic finds: the Baconsthorpe Page-Turner/Tweezers (Baconsthorpe, Norfolk, Mercia, Great Britain, archaeological dating: 700–800 CE) and the Sedgeford Runic Handle/Ladle (Sedgeford, Norfolk, Mercia, Great Britain, archaeological dating: 700–1000 CE). With the help of the formal description, it is possible to classify both new signs as very probably being runes. As such, they will be entered into the database, thus being searchable and analysable at the formal level, i.e. at the level of graph types and graph-type variants (see below, 2.2.3) as well as at the functional level.

### 2.2.2.1 Backbone of the Database: The *Find* Fields

The first prerequisite for the graphemic part of the database with regard to the comparative, comprehensive, and context-sensitive layout of our investigations was the systematization and digitalization of all “hard facts” about the runic inscriptions and the *Runica Manuscripta*. Hitherto, the runic finds have mostly been studied in different philological research traditions (Scandinavian studies, English studies, German studies), and also, with regard to the objects and their socio-historical context, by historical disciplines such as art history, archaeology or history of religion. This is the first time they have all been brought together as a digital and online accessible corpus, called the *find* fields (Figure 2.1), covering the following aspects of each monument:

1. Basic information has been collected, both from the relevant literature as well as through autopsies of our own, and by communicating with the respective institutions where the object is currently located. This comprises data relevant to identifying the runic monument, including the *find-spot* and the *object*, the *common names* of the runic find, and the *common abbreviations*. Furthermore, there is information on the *find year*, the *present location* of the inscribed object, the *state of preservation of object and inscription*, and the *inventory number or numbers*.
2. With regard to the inscription itself, and for a first overview, a *transliteration*, an *interpretation*, and a *German and English translation* are given. These are at present – from the perspective of the *RuneS* project and its investigations – “beta-data” only and serve mostly practical purposes. They represent “the state of the art”, the basic results so far achieved by runic research on the individual inscriptions. As our own planned *RuneS* research may lead to revised versions of these earlier transliterations and interpretations of a number of runic inscriptions, there will be additional versions of these data fields for internal use only. Once the graphemic and text-pragmatic investigations have been concluded, and consolidated, revised data are available, these data will replace the earlier “beta-data” and be made available online as well. This modification is an ongoing process, making the database a reflection of the research process.
3. Apart from these data, the *find* fields provide contextual information on the runic objects and inscriptions at several levels, covering the following areas: the *material*, and the *size and dimension*, i.e. the *measurements*, of the object; a *typological classification of the object*, e.g. tool, weapon etc.; information on the *archaeological or historic-cultural context* of the object; a (tentative) *dating* of both object and inscription (also in relation to context) – this is based mainly on archaeological suggestions; the *category of the inscription*, comprising classifications such as *runic*, *bi-scriptal*, *mixed*, or *coded* inscription; information on accompanying symbols, such as iconographic elements; an attribution of the inscription to a specific *rune row* (e.g., older *fupark*, younger *fupqrk/fupork* or Anglo-Frisian *fuporc*). All these data represent contextual areas, which may

trigger graphic or textual variation and, thus, be decisive for the graphemic and pragmatic evaluation and interpretation of the runic inscriptions.

4. In addition to this, GPS data for all geographical information (i.e. *find-spot*, *present location*, and the presumed place of origin, i.e. the *provenance* of object and inscription – the latter will be added in the course of the investigation), and images of the runic finds are provided.

The design of this first part of the database is influenced in many areas by, and has enormously profited from, already existing databases such as *Rundata*,<sup>4</sup> the *Danish online database*,<sup>5</sup> or the *Runenprojekt* database,<sup>6</sup> as these already present detailed and valuable information on the runic inscriptions of the respective corpora. This applies in particular to the inscriptions in the so-called younger *fuþarƿk/fuþork*, and to the fields of *transliteration*, *interpretation* and *English translation*. However, the *RuneS* database contains additional data in that it also includes the English, Frisian and South Germanic inscriptions as well as the *Runica Manuscripta*. The graphemic and pragmatic data, to be entered subsequently, will be unique to the *RuneS* database, going beyond the description of the object and the inscription by generating, as well as documenting, thematically-based research results.

The screenshot shows the 'Aspa stone' entry on the RuneS database. At the top, there is a navigation bar with 'Project', 'Find list', 'Default queries', 'Find map', and 'Advanced search'. Below this is a map showing the location of the stone in Aspa, Sweden. To the right of the map is a photograph of the stone. Below the map and photo are several tables of information:

Information on the object			
Class of object	stone	Material class	stone
Type of object	runestone	Material	fieldstone
Object	stone	Dimensions	185.62-38-
State	good	Completeness	yes
		Ext. dating	725-1100
		Method of dating	arch.
		Context	infrastructure: road, thingstead
		Find year	1667-?

Information on the inscription			
Runerow	younger fuþarƿk/fuþork	Additional markings / icons	yes
Category	run.	State	impaired: good
		Completeness	no

At the bottom right, there is a 'Place of storage' section with the text: 'Photographer: Bengt A Lundberg - 1985-06-19', 'Owner of rights: Riksantikvarieämbetet', and 'Type of rights: CC BY'.

Figure 2.1: Screenshot of the basic (= *find*) information on the Aspa stone (Sö 137), cf. [runesdb.eu/find-list/d/fa/q/////6/f/4782/] (for an overview of the runic objects of our corpus see [runesdb.eu/find-list])

4 [www.nordiska.uu.se/forskn/samnord.htm].

5 [www.runer.ku.dk].

6 [www.runenprojekt.uni-kiel.de].

### 2.2.3 Design of the Database – Step Two: The Graphemic Section and the Structure of the Database

The purpose of the graphemic analysis is to investigate the relation of sign and sound with regard to the system, as well as with respect to historical and regional variance. The assumption is that we are not dealing with a simple assignment of an individual sound to a sign, but rather with the two entities – sign and sound – each belonging to their own system, the graphemic and the phonemic system respectively, with a systematic functional relation holding between them.

Consequently, the graphemic analyses are divided into two working units: in the first step the graphs occurring in the inscriptions are described with regard to form. A typology of the runic graphs has been devised on the basis of graphic similarities. The second step focuses on the functional content of the signs. This approach makes it possible to answer various questions concerning the relation of the two systems, sound system and sign system, with regard to the so-called grapheme-phoneme correspondences in the runic script.<sup>7</sup> In addition, by including the runic separators and beginning and end marks, supra-segmental language and communication functions can be determined.<sup>8</sup> This approach also enables the detection of different types of graphic variants in the material, which can then be analysed regarding their distribution, e.g., in time (diachronically) and space (diatopically). As the formal characterization and typology of the graphs already implies a certain amount of generalization, leading to a first level of abstraction, this description cannot take place on the same data-level as the documentation of the inscription itself, i.e. not within the group of *find* fields, as these represent the level of the actual realizations of the runic signs (reflected in the database by a full-size image of the inscription and its signs)<sup>9</sup>.

Due to the fact that inscriptions in general consist of more than one graphic sign, all of which need to be formally classified in a different way, we also have a multi-relational connection between these data. On the other hand, the graphic similarities of two given runic signs in two different inscriptions – i.e., the one on object A, the other on object B – may be so close that they would have to be generalized as realizations of one and the same type of runic sign. This in turn means that the database needed to be conceived of as a relational database with a bilaterally multi-relational structure.

---

<sup>7</sup> An illustration of this approach to the formal description and functional interpretation of a recently discovered new rune in an inscription from Sedgeford can be found in Waxenberger, 2017.

<sup>8</sup> This is relevant for the subsequent text-linguistic and pragmatic issues, some of which are discussed in Zimmermann, 2017, along with an illustration of the text-linguistic and pragmatic approach applied to the Rø Stone.

<sup>9</sup> It should be pointed out that the process of negotiating copyright issues with various museums over the production and use of such snippets is highly time-consuming and at times very problematic.



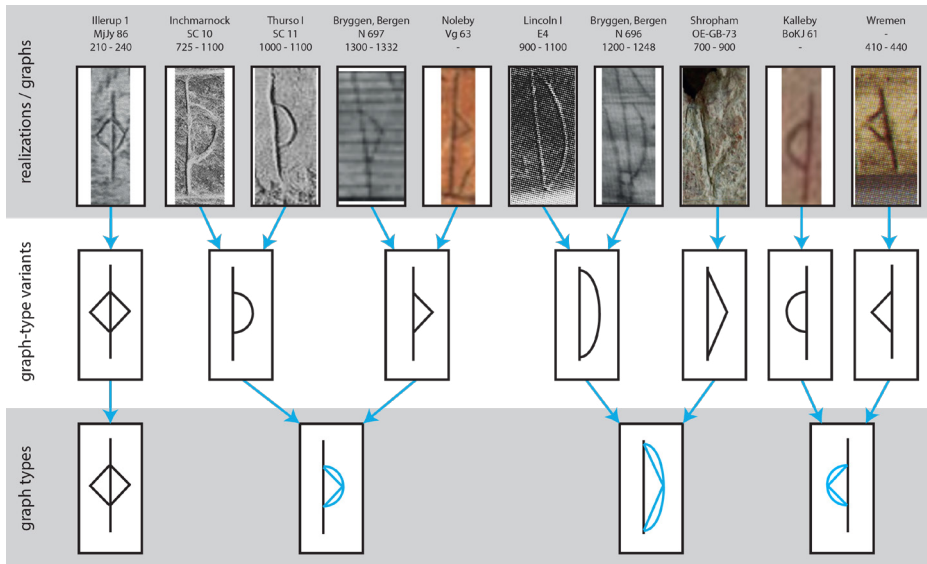
Due to the comprehensive structure of the corpus of investigation in the *RuneS* project and its context-sensitive approach, the description template for the assignment of the graph to a typology must take into account the respective realizations of the signs in the different rune rows (with regard to their potentially systematic functionalization), as well as graphic features that might be due to contextual and socio-historical factors only. The description template therefore needs to be designed in a very fine-grained way, e.g. concerning the position of the twigs and hooks on the stave as well as their specific form and technical execution. Thus, the typological characterization of the graphs in the database takes account of both micro- and macro-typographical features (Figure 2.2). Our typology is differentiated into:

1. *graph-type variants*, a lower level of abstraction where finer details in the execution of graphs are registered (Mårtensson, 2011, 115ff.), e.g., type of vertex (“crossing” or “with” resp. “without contact”) or “rounded” vs. “angular” form of compositional elements, etc. Even at this lowest level of formal characterization, the signs described are not confined to a single occurrence in one inscription only. This means it is possible to set the formal characterization of a variant in relation to several inscriptions and the realizations found there.
2. *graph types*, a level of abstraction where the graphic variation taken into account concerns the “basic shape of the graphs and their distinctive [formal] features” (Mårtensson, 2011, 113ff.), i.e.: number of the elements stave, twig, hook and dot and their position in relation to each other, as well as the elements involved in each vertex.
3. The individual realizations are included for the sake of illustration in the respective graph-type variant table as *snippets*.

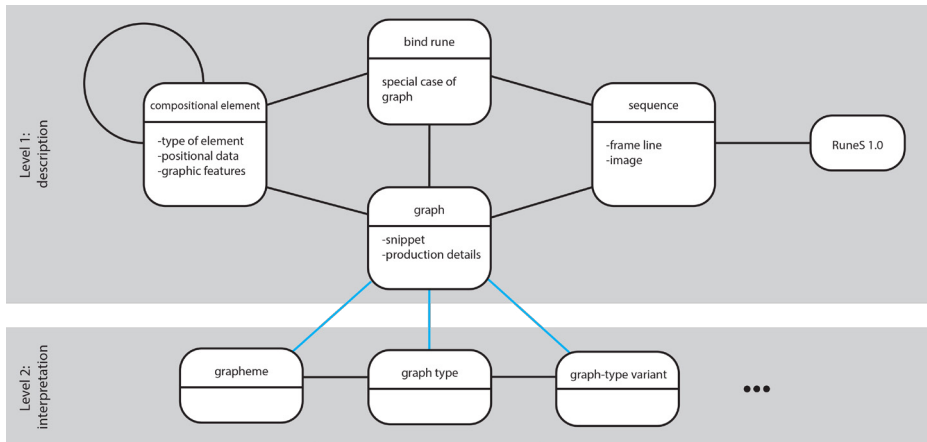
We thus have two levels of increasing formal abstraction in addition to the snippet of the individual graphs (visually, the two abstraction levels are presented by graphic depictions of the respective “types” = depiction of the typologically indicative features).

The functional analysis, i.e., the so-called grapheme analysis, represents a further level of abstraction. Thus, additional individual datasets are required to allow for linking different sound-related and non-sound-related functions to the formal realizations and the various formal types of the graphic variants.

On the technical side these layers of abstraction are implemented using a hierarchical structure of entity types. Thus, the relationships between the graphs, individual compositional elements and the interpretation/categorization of a graph are distributed over different parts of the database. The collection and generation of graphemic data for the database is performed in two steps, also reflected in the database structure (Figure 2.3).



**Figure 2.2:** Random sample of snippets of the  $\beta$ -rune with the assigned graph-type variants and graph types



**Figure 2.3:** Highly simplified structure of the graphemic section of the database

On the first level, the compositional elements are described individually. For each type of element (staves, twigs, hooks and dots – and the individual vertices) coordinates and details concerning their graphic form are stored in a database table. A hierarchical tree structure of these elements is employed to allow for the automatic extraction of information on the relative position and relationships between individual elements. This tree structure also enables the efficient categorization of special cases like bind

runes, where several graphs are combined into a single graph. Furthermore, data concerning the graph as a whole are collected. Most importantly, an interpretation of the graph is given, including its mapping onto a graph-type variant, a graph type and other features (cf. blue lines in Figure 2.3).

On the second level, data concerning the historical relationships between graph types, graph-type variants, graphemes etc. are stored in the database. These relationships are populated using the graph interpretations given for individual graphs and are set to emerge automatically while data concerning individual graphs are collected. This structure represents the actual use of graphs and their functions, making it a very useful research tool which will be available to the public on the *RuneS* website.

Using the graphemic data collected in this step, a number of research tools can be offered to our users, allowing for a range of sophisticated research questions. For example, when combining these data with the *find* data collected in *RuneS 1.0*, historical and geographical distributions of the uses and forms of graph-type variants and graph types can be automatically visualized.

## 2.2.4 Design of the Database – Step Three: The Bilingual Layout

### 2.2.4.1 Bilingual Terminology: Choices

In the context of establishing the graph-typological description templates, it is also necessary to decide on terminology. Due to the bilingual character of the database, this means not only deciding on a single term for a certain phenomenon in each language, possibly from a panoply of already existing usages, but to decide on twin terms in both languages for each and every feature to be entered, thereby ensuring identical search potential and identical search results in the two language versions. Each term pair therefore needs to be discussed with regard to its internal compatibility.

This has sometimes led to the rejection of established terms, such as the terms *Lesung* vs. *reading* (see above, 2.2.2.1, the *Find* Fields), as a survey of selected research sources revealed that the latter, the English term *reading*, has been used in a wider sense than the German term, including both transliteration and interpretation. This is not compatible with the database structure and the terms have therefore not been employed here. Instead, we are using the English-German set: *transliteration/Transliteration*.

In the context of the graph-typological description, it became necessary to narrow down and systematize the existing terms for the elements a runic sign may be composed of. In English, these were (*main*) *stave/staff*; (*side*) *twig, branch*; *hook, crook, chevron, angle, pocket*; in German we found (*Haupt-*)*Stab*; *Zweig*; *Haken, Buckel*. Our selection, to be implemented for the first time systematically in the graph-typological entries into the database, is as follows: *Stab* – *stave*; *Zweig* – *twig*; *Haken* – *hook*; *Punkt* – *dot*.

In this way, the selection and refinement of terminology made necessary through the database requirements will hopefully also lead us to greater precision at the content level, while at the same time instigating reflection on the suitability of established terminology.

#### 2.2.4.2 Bilingual Terminology: Technical Aspects

The support for multiple languages has to be deeply integrated into the database design. Our database uses two different approaches for different kinds of fields, where the nature of the data is the differentiating factor (Figure 2.4).

run_constants				run_find				
id	group	trans_de	trans_en	id	obj_complete	material	is_trans_de	is_trans_en
42	obj_complete	„ja“	„yes“	1	42	45	„ich, Widuh...“	„I, Widuh...“
43	obj_complete	„nein“	„no“	2	42	45	„Alugod“	„Alugod“
44	material	„Stein“	„stone“	3	43	46	„Ich, Unwöd...“	„I, Unwöd...“
45	material	„Metall“	„metal“	4	43	44	„Sie kämpften...“	„They fought...“

Figure 2.4: Simplified view of bilingual data in the database

For most fields in our database the number of possible values is finite, e.g. the field *completeness of the object* can only carry the values “yes” (“ja”) or “no” (“nein”). These constants are stored in a separate table, *run\_constants*, with columns for each language. The database fields in question contain pointers to the corresponding values in *run\_constants* that are substituted for their translations when data is displayed. In order to differentiate which constants belong to which fields, a grouping column has been introduced. Note that all this happens in the back-end and is not visible during data entry or on our website.

Other database fields like the *translation of the inscription* may contain very specific data and cannot be included in *run\_constants*. For these types of fields, we have employed multiple fields in the same interface (a German one and an English one).

An ideal solution for supporting additional languages would be a table with the columns *id*, *reference*, *group*, *language* and *translation*, where the column *reference* is used to store the pointers employed throughout the database and *language* contains a unique identifier for the language of the *translation*. This way the database would support an infinite number of languages without the introduction of additional columns. However, since our database is not likely to support additional languages, we decided that the computational overhead of this approach would outweigh the benefits.

### 2.2.5 Design of the Database – Step Four: Data Mask for the Input of Graphic and Graphemic Data

The graphemic data collected in the database are of an extremely visual nature since elements are described according to their size and relation to each other. A lot of these data may be computed automatically once the exact layout of the individual compositional elements is known. Thus, an interactive mask has been programmed that focuses on the visual aspects of the compositional elements of the graphs rather than the specific data stored in the database. Users can place compositional elements directly onto a reference image (if one is available) and specify their positions by simply moving them around (Figure 2.5). The system automatically computes relevant research data from the coordinates of the graphs and their compositional elements and offers a range of options to further specify the nature of these elements (e.g., *tools used for production* or *the sequence of production*). Data entry is performed in three distinct steps:

1. The position of the *frame line* defining the upper and lower boundaries of the sequence of graphs is collected.
2. The *compositional elements* of the graphs are described individually. In this step, coordinates and details concerning the form of staves, twigs, hooks and dots are collected. Elements are organized in a hierarchical tree structure, allowing us to extract information on the relative position and relationships between individual elements automatically.
3. Data concerning the graph as a whole are collected. Most importantly, an interpretation of the graph is given in this step, i.e. its mapping onto a graph-type, a graph-type variant and other features. These mappings are automatically informed by previous mappings assigned to visually similar graphs: when choosing which graph-type the current graph belongs to, the mask automatically suggests graph-types that are visually similar to the graph in question. This way the user's navigation of a large network of graph-types is assisted and the efficiency of data entry is improved.

As mentioned in section 2.2.3, a structure representing the historical relationships between graph interpretations is set to emerge automatically from the data given in step three. However, this is set to happen under human supervision, and an interface for the analysis and regulation of these data will be created. The data collected this way represent the actual historical use of graphs and their functions, making them a very useful research tool that will be available to the public on the *RuneS* website.

In addition to automatically handling the storage of the positional and relational data in the database, the mask also extracts snippet images for each individual graph (if a source image is available). The use of a visual and guided tool like this has the additional advantage that errors in the data are immediately visible, while they would potentially remain hidden and obscured if these data were collected solely in text form.



Figure 2.5: The graphemic data input mask

## 2.3 Concluding Remarks

To sum up, the *RuneS* database will ultimately consist of a documentation of the runic finds, a graph-typological (i.e. formal) as well as a graphemic (i.e. functional) analysis of all runic signs, and a text-linguistic and pragmatic description and analysis of the complete inscription in context. It will display the research results of the *RuneS* project and enable users of the online version to combine the provided data according to their own research objectives.

Naturally, the development and implementation of such a complex digital tool has not been without its specific problems. One of the recurring problems during the implementation of the first two parts of the database was the reduction of complexity engendered by a digital database. In some cases, this initial problem proved to be a fruitful catalyst for reaching new clarity, e.g. in the development of a new, joint bilingual terminology for the labels of the individual fields or the options within the fields, or in coming to more theoretical and methodological accuracy with regard to the transliteration system of the inscriptions. This meant scrutinizing traditional, runological terminologies in both German and English, and establishing a common usage within the project.

However, the structure of different fields with clearly defined options, while enabling and facilitating research by the ensuing searchability, may lead to the obfuscation of open questions. In order to make transparent such open issues while preserving the searchability of the database, different solutions were developed. Where

a set of data did not fit clearly into any one of the categories evolved from the bulk of the material, either due to the state of research or the nature of the object to be categorized, this was marked by giving it a “dual label”, i.e. a dual categorization. This was the case, for example, with the classification of an inscription as “older fuþark” or “younger fuþark/fuþork”. The dual value “older fuþark/younger fuþark” was integrated into the list of options of the data field “rune row”, e.g. for the Lousgård bead or the Roes stone. An open commentary field reflects the state of the art with regard to the issues under debate. Here, the user may also find differing interpretations and datings, etc.

The next step, after the completion of the graphemic part of the database, will be the development of the text-linguistic and pragmatic part (*RuneS 3.0*). Future directions also include linking with the respective data sets of other digital projects that are thematically relevant: digital versions of runic editions (e.g., *Digitala Sveriges runinskrifter*),<sup>10</sup> online dictionaries such as the *Dictionary of Old English*,<sup>11</sup> or archaeological databases such as the *Portable Antiquities Scheme*.<sup>12</sup>

We hope very much that we are, in this way, in the process of building a database that will not only help us in conducting our own *RuneS* research, but also serve as a digital information platform and a search tool for all colleagues interested in runes, as well as runic and other forms of epigraphic writing.

## Bibliography

- Koch, P. & Oesterreicher, W. (1985). Sprache der Nähe – Sprache der Distanz. Mündlichkeit und Schriftlichkeit im Spannungsfeld von Sprachtheorie und Sprachgeschichte. *Romanisches Jahrbuch*, 36, 15–43.
- Mårtensson, L. (2011). *Studier i AM 557 4to. Kodikologisk, grafonomisk och ortografisk undersökning av en isländsk sammelhandskrift från 1400-talet*. Reykjavík: Stofnun Árna Magnússonar í íslenskum fræðum.
- Page, R.I. (1973). *An Introduction to English Runes*. London: Methuen.
- Waxenberger, G. (2000). The Inscription on the Gandersheim Casket and the Runes in the Old English Runes corpus (Epigraphical Material). In R. Marth (Ed.), *Das Gandersheimer Runenkästchen. Internationales Kolloquium Braunschweig, 24.-26. März 1999* (pp. 91–104). Braunschweig: Limbach Druck und Verlag.
- Waxenberger, G. (2017). A New Character on the Sedgford Runic Handle/Ladle: Sound Value Wanted. *Anglia*, 135(4), 627–640.
- Zimmermann, C. (2017). Interdisziplinäre Interpretation: Theoretische Grundlagen und methodische Ansätze. In J. Krüger, V. Busch, K. Seidel, C. Zimmermann, & U. Zimmermann (Eds.), *Die Faszination des Verborgenen und seine Entschlüsselung: Rāði sār kunni. Beiträge zur Runologie, skandinavistischen Mediävistik und germanischen Sprachwissenschaft* (pp. 429–448). Berlin/Boston: De Gruyter.

<sup>10</sup> [<https://www.raa.se/kulturarv/runor-och-runstenar/digitala-sveriges-runinskrifter/>].

<sup>11</sup> [<https://www.doe.utoronto.ca/>].

<sup>12</sup> [<https://finds.org.uk/>].

María José Estarán, Francisco Beltrán, Eduardo Orduña and Joaquín Gorrochategui

### **3 Hesperia, a Database for Palaeohispanic Languages; and AELAW, a Database for the Ancient European Languages and Writings. Challenges, Solutions, Prospects**

**Abstract:** *Hesperia. Banco de datos de lenguas paleohispánicas* and *AELAW. Ancient European Languages and Writings* are two narrowly linked projects whose common feature is their general aim: cataloguing the documents written in the ancient languages of Europe (8<sup>th</sup> cent. BCE–5<sup>th</sup> cent. CE) excluding Latin, Greek, and Phoenician. Although both projects are closely linked, BDHesp has a track record of twenty years, while AELAW has been active for only two and a half years. In this paper, where we have especially focused on BDHesp, we summarize the problems that arose during the encoding of Palaeohispanic languages, written in multiple writing systems and their variants, and the solutions addressed. We also present the promising tools that have been developed in BDHesp to make significant progress in our understanding of Palaeohispanic languages and writings. Lastly, we introduce AELAW network and its two databases, its aims and what we intend to accomplish in the future.

**Keywords:** Palaeohispanic languages, Palaeohispanic writings, ancient languages of Europe, partially deciphered script, digital epigraphy

---

**María José Estarán**, University of Nottingham, LatinNow Project (ERC-2016-STG 715626)

**Francisco Beltrán**, Universidad de Zaragoza

**Eduardo Orduña**, IES Pont de Suert

**Joaquín Gorrochategui**, Universidad del País Vasco/Euskal Herriko Unibertsitatea, J. Gorrochategui is PI of the project “Hesperia: lenguas, epigrafía y onomástica paleohispánica”, funded by MINECO/FEDER (FFI2015-63981-C3). The rest of the authors of this paper are active members of this project.





### 3.1 Introduction to BDHesp and AELAW Databases

*Hesperia. Banco de datos de lenguas paleohispánicas*<sup>1</sup> and AELAW. *Ancient European Languages and Writings*<sup>2</sup> are two narrowly linked projects whose common feature is their general aim: cataloguing the documents written in the ancient languages of Europe (8<sup>th</sup> cent. BCE–5<sup>th</sup> cent. CE) excluding Latin, Greek, and Phoenician.

The purpose of *Hesperia. Banco de datos de lenguas paleohispánicas* (henceforth BDHesp) is to collect the inscriptions written in any of the pre-Roman languages known in Hispania and Southeastern Gaul, including coin legends. Its distinctive feature is that it is not a mere compilation of epigraphs (i.e. a *sylloge* or an *editio minor* of the texts known so far); on the contrary, it meets the criteria of a genuine *editio maior*, where every text has been analysed accurately and every file is provided with a critical apparatus of the text, as well as with pictures or drawings (Luján, 2005; Orduña, Luján & Estarán, 2009; Orduña & Luján, 2014; Orduña & Luján, forthcoming).

This project began in 1997 and it is currently being developed by four teams based in the Universitat de Barcelona (UB), Universidad Complutense de Madrid (UCM), Universidad de Zaragoza (UZ) and Universidad del País Vasco / Euskal Herriko Unibertsitatea (UPV/EHU), thanks to the funding of the *Plan nacional de I+D+i*, sponsored at present by the Spanish Ministerio de Economía y Competitividad. The project was initiated under the direction of Javier de Hoz (UCM) and it is presently coordinated by Joaquín Gorrochategui (UPV/EHU). J. Velaza (UB), E. Luján (UCM) and F. Beltrán (UZ) are the individuals responsible for the rest of the teams.

As for AELAW, it is focused on the creation of a network of researchers working on the European languages spoken and written in Antiquity, excepting Latin, Greek and Phoenician. Its final goal is to lay the foundations of a databank that could group every inscription written in one of these ancient European languages. Its medium-term partial goals are 1) to create a census of languages; 2) to create a census of inscriptions; 3) to fix the criteria for the digital edition of inscriptions; 4) to define the technical features of the future Databank. The AELAW network was born in 2015 thanks to a European *Cooperation in Science and Technology Action* (COST IS1407). This action will last until 2019.

The network, whose chair is F. Beltrán (UZ), is currently composed of researchers working for 29 institutions based in 13 countries. AELAW emerged as an initiative of the Spanish researchers belonging to the Hesperia project with the purpose of providing the ancient European epigraphic ensembles with a tool, which could be similar to BDHesp. As a consequence, both projects are closely linked. However, we would like to underline that BDHesp has a track record of twenty years, while AELAW

---

1 [<http://hesperia.ucm.es/>].

2 [<http://aelaw.unizar.es/>].

has been active for just two and a half years, hence the presentation of each database in this paper is clearly unbalanced towards the first project.

### 3.2 Palaeohispanic Languages and Writings

The Iberian Peninsula is a region with a high linguistic heterogeneity where three colonial languages (Phoenician, Greek and Latin) and five vernacular languages belonging to different linguistic groups are recognised. More than two thousand inscriptions written in these local languages and writings, dating from the 7<sup>th</sup> cent. BCE to 1<sup>st</sup>/2<sup>nd</sup> cent. CE (when they were substituted by the Latin language and alphabet) have been discovered so far. The four languages epigraphically recorded are, in a diachronic order: the so-called Southwestern language (or “Tartessian language”), Iberian, Celtiberian and Lusitanian. To these should be added the Vasconic language, known by onomastics and possibly by certain short texts, although it is still a controversial question.

The Iberian language is also recognised in Southeastern France (west of the Hérault river). Aquitanian, a language that was closely related to Vasconic, and is known only through some personal and god names, was spoken on the other side of Pyrenees. The Vasconic-Aquitanian remains are clearly linked to the currently spoken Basque language, albeit Ancient Vasconic is better attested in Navarre and the northern territory of Zaragoza (Aragon) than in the area corresponding today to the Basque Country, where the epigraphic records are mainly related to the Celtic languages.

The Celtiberian language belongs to the Celtic branch, such as Gaulish and Lepontic in Antiquity, or other currently spoken languages such as Brittonic, Gallic or Irish. The Lusitanian language is clearly an Indo-European language, although there is not yet consensus on its belonging to the Celtic branch, since Lusitanian presents some characteristics that differ from the Celtic features: the Lusitanian inscriptions retain Indo-European \*p-. The classification of the so-called “Southwestern language” poses even more problems, since its writing system is only partially deciphered. Some researchers consider that it is also a Celtic language, although it is a minority opinion. As for the Iberian language, researchers have been able to determine that it seems an agglutinative language. It remains unclassified, without known parallels, although it presents some similarities with the Vasconic group that are still insufficient to confirm a direct connection between both languages.

The texts that were written in these languages mainly used a writing system called “Palaeohispanic”, which originated in the Iberian Peninsula, whose most distinctive feature is the use of both alphabetic graphemes (for vowels, sonants and sibilants) and syllabic graphemes (unvoiced and voiced plosives). At least four variants (possibly five) of this “semi-syllabic” writing system have been identified: 1) The variant used for the inscriptions written in the “Southwestern” language; 2) and 3) The variants

of the Iberian-speaking region, along the Mediterranean coast and its inland, between Southern France and Almeria; 4) The Celtiberian variant, spread along the Sistema Iberico (an inner mountain chain); 5) and maybe a “Vasconic” variant in the Northwestern Middle Ebro Valley, where these Vasconic speaking peoples and other related peoples were settled. Besides, the Iberian language was written in a variant of the Greek Ionic alphabet and, exceptionally, in the Latin alphabet. The Latin alphabet was often used, in turn, for transcribing the Celtiberian language (with some minor modifications) in an advanced stage of Romanization. The Latin alphabet is also the writing system of the scarce Lusitanian texts, without exceptions.

These four linguistic groups cover the Southern half of Hispania and its Far East. On the contrary, the West remained illiterate until the Roman conquest (late 1<sup>st</sup> cent. BCE), where no vernacular language is occurring in any inscription, since their texts were written in the Latin language from the beginning.

The Southwestern language is evidenced between the 7<sup>th</sup> and 4<sup>th</sup> centuries BCE on *instrumenta*, but above all on stone: funerary texts with a striking helicoidal layout were inscribed on *circa* two hundred stones. The variant of the Palaeohispanic writing system used there has not been completely deciphered up to date and, as a consequence, the linguistic classification of this poorly known language is under discussion.

The Iberian language is the best evidenced of all the Palaeohispanic languages. More than two thousand inscriptions written in this language are dated between the 5<sup>th</sup> century and the 1<sup>st</sup> century BCE (some epigraphs could even be dated in the 1<sup>st</sup>–2<sup>nd</sup> centuries CE). The oldest inscriptions are written in a variant of the Greek alphabet (in a restricted region near Alicante) or in a variant of the Palaeohispanic writing system. Short texts on *instrumentum*, and longer texts related to trade or economic activities on lead tablets are the documents one can find in the earliest stages of the Iberian epigraphy. In tandem with the Roman conquest, literacy spread inland from the 2<sup>nd</sup> century BCE onwards. From that moment on, we move to an intensification and diversification of the epigraphic habit: monumental inscriptions, aimed to be publicly displayed and contemplated, and funerary steles and slabs are the most remarkable novelties; but important changes in coin legends and mosaic inscriptions occurred as well. The lack of linguistic parallels for the Iberian language makes this language very difficult to understand. Only personal names have been identified, and the sense of some words has been perceived in only a tentative way.

Simultaneously, from the 2<sup>nd</sup> century BCE on, literacy spread to the Celtiberian area, where some hundreds of inscriptions have been collected, mostly on *instrumentum*, albeit scarcely more than a dozen on monuments on stone have been found, as well as twenty instances of graffiti inscribed on a rock sanctuary, approximately forty *tesserae hospitales* on little bronze objects and around ten inscriptions, some of them extraordinarily long, on bronze plaques and tablets. Although the Celtiberian language cannot yet be translated, the linguistic comparison with other Celtic and

Indo-European languages allows an understanding of its morphology and syntax and, therefore, of some words and word sequences.

Lastly, the Lusitanian language is only evidenced in half a dozen stone and rock inscriptions of a religious nature, where some theonyms and references to animal sacrifices have been identified. All Lusitanian texts are written in the Latin alphabet. Apart from this small ensemble of inscriptions, a group of altars inscribed with the Latin language from the Lusitanian region, bear religious dedications to local gods whose theonyms show Lusitanian morphological traces.

The reference work for Palaeohispanic inscriptions are the four volumes of *Monumenta Linguarum Hispanicarum*, published by Jürgen Untermann since 1975, that BDHesp intends to update.

### 3.3 BDHesp (Banco de Datos de Lenguas Paleohispánicas Hesperia)

*Hesperia*, the Databank for Palaeohispanic Languages (Figure 3.1) is based on Untermann's *MLH* and Wodtko 2000 and, following the lines of this corpus, the epigraphic material is organized territorially. This is structured according to the current Spanish and Portuguese provinces and the French *départements*.

The implementation of this project has been possible thanks to Eduardo Orduña. He has built the Databank on LAMP, the software bundle consisting of the software operating system Linux, the web server Apache, the relational database management system MySQL and the programming language PHP, all them leading representatives of the free software and of the open source code, as well as MapServer, developed by the University of Minnesota, which has been used for the generation of the maps.

The aim of *Hesperia* is, with the help of computational resources to create a linguistic and epigraphic database that allows us to develop our precarious knowledge of Palaeohispanic languages and writings. This databank facilitates compilation of all the published Palaeohispanic inscriptions with a complete set of information (the identification of each inscription and its text; epigraphic, linguistic and archaeological commentaries; bibliographic references and pictures), as well as adding new files and improving those previously published. The fact it is not a simple data collection, but a critically edited file, is what gives BDHesp a relevant scientific value.

The image shows the home page of the Hesperia website. At the top left is the logo for 'HESPERIA BANCO DE DATOS DE LENGUAS PALEOHISPÁNICAS'. To the right of the logo is a search bar with dropdown menus for 'Base Epigráfica', 'Numismática', 'Onomástica', 'Bibliografía', and 'Mapas'. Below the search bar is a navigation menu with the following items: 'PRESENTACIÓN', 'Presentación', 'El proyecto Hesperia', 'Proyecto actual', 'Equipo', 'Proyectos anteriores', 'DOCUMENTACIÓN', 'BANCO DE DATOS', 'ENLACES', 'NOTICIAS', and 'CONTACTO'. The main content area is titled 'Mapa de las monedas e inscripciones paleohispánicas, clasificadas según zonas epigráficas y lingüísticas.' and features a map of the Iberian Peninsula with various colored markers and letters (A-J) indicating different linguistic zones. A small image of a coin is labeled 'A'. At the bottom, there is a logo for 'GOBIERNO DE ESPAÑA' and 'MINISTERIO DE CULTURA Y PATRIMONIO'. The footer contains the following text: 'KUTIM - 08 - 2018 - AAA - Diseño Web HTTP://AMIBARROMANO.ES. Diseño de la base de datos: RICARDO OJEDA ANAS. © 2008 DEPARTAMENT DE FILOLOGIA GREGA I LINGÜÍSTICA INDIEUROPEA - UNIVERSITAT COMPLUTENSE DE MADRID. LICENCIA DE USO: http://www.derecho4u.com Última actualización del banco de datos: 30-11-2017.'

Figure 3.1: Home page of Hesperia

### 3.3.1 Developing BDHesp: From an Epigraphic Database to a Databank of Palaeohispanic Languages

The BDHesp coordinators decided to create two more databases besides the epigraphic one: a database for coin legends and another one for bibliographic references. The numismatic database (Estarán & Beltrán, 2015) was clearly inspired by Untermann's *MLH*, where epigraphy on coins is collected in an independent volume. The information included in this database doesn't cross automatically with the epigraphic database, while the files in the bibliographic database do. This second database contains all the bibliographic references mentioned both in the epigraphic and in the numismatic database. As a consequence, the original epigraphic database becomes a databank. As BDHesp progressed, more tools were created and linked to the numismatic and epigraphic databases: a map generator and a search engine. These were improved, as new needs arose in the creation of files. At present, the BDHesp team continues to develop new tools and databases. Indeed, a brand new database has been recently opened within this databank to collect onomastics (Vallejo, 2016), since personal names and theonyms play a fundamental role for research on the ancient languages of certain areas where texts written in the vernacular languages are absent (Gorrochategui & Vallejo, forthcoming). Work-in-progress is being carried

out in another database devoted to the lexicon, which will be accessible in the very near future.

Likewise, the database of the ENCEOM project (ENCEOM, *El nacimiento de las culturas epigráficas del Occidente Mediterráneo*, Ministerio de Economía y Competitividad, PI: F. Beltrán Lloris) has been recently prepared to be incorporated into BDHesp. Although this project is not directly related to Hesperia, its files were designed to be compatible with BDHesp from the very beginning. This database currently includes more than 750 files of publicly displayed inscriptions of the most relevant epigraphic cultures of the Roman West (Iberian, Celtiberian, Lusitanian, Gaulish, Oscan, Umbrian, Phoenician / Punic and some Etruscan inscriptions as well).

The multidisciplinary nature of the team of project Hesperia has been essential for developing very complete files, containing comments on the epigraphic material, the archaeological context, the linguistic exploitation of the texts, etc. All this information can be easily found thanks to the search engine.

### 3.3.2 Challenges Arising from the Digitalization of Palaeohispanic Epigraphy and Solutions Addressed in BDHesp

The digitalization of Palaeohispanic epigraphy has posed some challenges related to codification and computational lexicography.

- Codification. The main problems regarding codification that have been faced are, on the one hand, processing texts written in different writing systems and, on the other hand, the existence of certain Palaeohispanic graphemes, on whose phonetic content there is not yet consensus among the researchers. The first one has been relatively solved thanks to the transcription of the Palaeohispanic texts in the Latin alphabet within the files.<sup>3</sup> E. Orduña created buttons with diacritic symbols, or for introducing bold and italic letters, in order to facilitate the introduction of Palaeohispanic texts without having to deal with codes, which might have caused several problems if done incorrectly. Regarding the second problem, E. Orduña proposed the option “Personalizar transcripción” (“Customize transcription”), which gives the user the possibility of choosing the phonetic value assigned to every doubtful grapheme. A specific search engine based on this system has been implemented for the texts of the inscriptions written both in the Southern Iberian script (a variant of the Palaeohispanic Iberian script) and in the Southwestern writing system, which are only partially

---

<sup>3</sup> If the potential user is interested in knowing more about Palaeohispanic writing systems, he or she can visit [<http://hesperia.ucm.es/escrituras.php>], where some explanations and tables with graphemes and allographs have been uploaded.

deciphered. Its strength lies in the possibility of assigning different values to every sign, permitting the different reading options to be seen immediately. Additionally, the user can see the undeciphered graphemes as images. However, the best-known variants of the Palaeohispanic writing system (Northwestern Iberian and Celtiberian variants) present specific problems of codification, namely the existence of different transcription systems in the current research. This problem affects mainly inscriptions distinguishing the marked and unmarked syllabograms that might correspond to a distinction of voiced and voiceless plosives respectively. Some researchers transcribe them as such (e.g. **ta/da, ka/ga**); some others prefer a more restricted system that only reflects the marked or unmarked nature of the Palaeohispanic grapheme (**tá/ta, ká/ka**). Through the internal use of regular expressions, it has been possible to develop a system that allows the user to choose the transcription system that best suits his or her needs. Lastly, a problem concerning every transcription system is the use of underdots, or underlining, to mark a doubtful or incomplete reading, which means an added difficulty for the search engine. This problem has been solved using Unicode diacritics to transcribe these signs, in order to benefit from the power of the system of regular expressions of the PHP language, which permits them to be ignored in the searches.

- Access to information. BDHesp has been designed with the aim of facilitating the access to the huge amount of data contained in it, which has been solved in diverse ways: 1) the user can get access to the files not only from the search engine, but also from the map server. The maps contain clickable marked places that connect automatically to the corresponding epigraphic file; 2) once the user has filtered the information with his or her desired criteria through the search engine, the user can choose the layout of the results (like a list or like the pages of a book), in order to provide comfortable reading; 3) if the user is looking for certain regular expression, which is especially useful for determining patterns in Palaeohispanic texts, he or she can introduce the desired expression in the search engine. It will provide a complete list that may include eventual variants. These possibilities make the BDHesp search engine an indispensable tool to make significant progress in the deciphering of Palaeohispanic texts, since it offers an easy access to data that otherwise would be very tiresome to obtain: the reading variants of the search results appear as bubbles on the selected reading when the cursor is hovered over them; in the same way, bibliographical references appear on the abbreviated ones. Lastly, BDHesp developers have not only considered the screen layout, but also the printed layout: it is possible to generate PDF files with all the information, or the data the user has previously selected, of one file or a group of them, including pictures and drawings of the inscriptions.
- Computational lexicography. Each database in BDHesp has different aims, so BDHesp developers reflected deeply on the special needs of each one before achieving the final design, and therefore on the units in which these databases

were going to be structured. The dominant criterion for structuring each unit was, in all cases, the ease of reference online.

For example, the unit is the mint (the city that issued the coins) in the numismatic database, thus files for coin legends are grouped in their respective mint files. If the unit were every different legend, the searches would have been less straightforward for the user (nonetheless, this database has two combined search engines: one for mints and another for coin legends, in case the user should be interested in a particular coin legend). On the contrary, the lexicographic unit of the epigraphic database is not the archaeological site or the ancient city where the inscriptions were found, but the inscription itself. Of course, inscriptions are geographically grouped; but there is not a specific file for each site or place. On the other hand, the unit of the onomastic database is obviously different from the other two databases: each personal name, theonym or toponym is the unit of a file. This selection was fundamental both to know their frequency and the cartographic distribution. Similarly, in the lexicon database, each lexical element must be isolated so as to be individually studied.

In this sense, an additional problem has emerged when linking the epigraphic database with the lexicon database: the identification of “words”. In essence, it is already solved, although it is not yet publicly accessible. For instance, we cannot yet identify “words” in Iberian with certainty, given our precarious knowledge of this language. That is why each entry of the BDHesp lexicon corresponds with the segments that were separated with interpuncts by the Iberians themselves. The programme uses these signs to internally convert the text of the inscription in an array with each segment, and, after that, it executes a loop comparing each element of the array with the entries of the lexicon. Then it generates a new version of the text on the screen, where every word appears like a link to its corresponding entry of the lexicon. The use of regular expressions in the comparison even allows the creation of links to non-exact corresponding entries, ignoring lost signs or problems of transcription, for example.

- Small-scale geographic view. We have already mentioned the possibility of dynamically generating location maps of inscriptions. The existing possibility of loading layers of external servers, like Google Maps (with satellite view) or local layers (like georeferenced maps), allows us to foresee future challenges: a collaboration with archaeologists could provide precise geographic coordinates for the location of findings in a site, so that we could visualize the distribution of the inscriptions on the satellite photograph or on the georeferenced plan of the site.
- Interoperability. BDHesp has not yet taken the leap to the compatibility with other epigraphic databases, probably because no other database is thematically related (only with the future database AELAW, see below); and a need of associating with a thematically unrelated database has not arisen, since, for example, the mapping software is already incorporated in BDHesp.



In sum, BDHesp could be considered as the indispensable tool for researchers in Palaeohispanic languages and cultures. Thanks to its computational resources (search engine, mapping software, the possibility for the user of reading simultaneously the official reading and its variants, or of choosing the phonetic values for the doubtful graphemes, etc.), the research is going to progress profoundly in our knowledge of the Palaeohispanic languages and writings.

### 3.4 AELAW

The concept of AELAW is clearly different from that of Hesperia. The COST Action *Ancient European Languages and Writings* began in 2015 and, as has been underlined before, it is inspired by BDHesp to a large extent. The main aim of this action, funded by the European Union through the programme European Cooperation in Science and Technology (COST), is to create a network of researchers working on ancient European languages and writings through the establishment of links between universities and research centres. This network will overcome the existing fragmentation among the researchers of the different Palaeo-European epigraphic cultures.

This network must generate links that ease the cooperation, the exchange of experiences and the sharing of advances made in the research on each corpus language in order to find solutions to the various problems each region poses. The training of early-stage researchers through short-term scientific missions, training schools, workshops and conferences, is considered particularly relevant. Additionally, we intend to establish the criteria for critical editions online and to develop a databank that will contain all the Palaeo-European inscriptions.

The AELAW network promotes multiple scientific activities and meetings and publications, among which the collection of AELAW *Booklets* stands out. These booklets provide accurate and attractive introductions to the epigraphic production of each fragmentary, but evidenced language (Beltrán & Jordán, 2017a, 2017b; Salomon, 2017a, 2017b; Velaza & Moncunill, 2017a, 2017b; Wodtko, 2017a, 2017b).

The image shows the home page of the AELAW Database. At the top, there is a dark blue header with the AELAW logo and navigation links: EVENTS, PEOPLE, ITEM, WORKING GROUPS, PUBLICATIONS, DATABASE, and BLOG. Below the header, there is a breadcrumb trail: HOME / DATABASE / languages. The main content area is titled "DATABASE" and has three tabs: LANGUAGES (selected), INSCRIPTIONS, and BIBLIOGRAPHY. The "LANGUAGES SEARCH" section contains several filters: LANGUAGE NAME, NUMBER OF INSCRIPTIONS, LANGUAGE FAMILY, LOCALIZATION, CHRONOLOGY (with FROM and TO dropdowns), TYPOLOGY OF OBJECTS, MAIN TYPES OF TEXTS, and WRITING SYSTEMS USED. A "SEARCH" button is located at the bottom right of this section. Below the search filters is a "SOURCES (BIBLIO)" section with six input fields: PRINTED CORPORA, DIGITAL EDITIONS, JOURNALS, CONFERENCES, OTHER PRINTED, and UNPUBLISHED SOURCES. A "SEARCH" button is also present at the bottom right of this section.

Figure 3.2: Home page of the AELAW Database<sup>4</sup>

### 3.4.1 Developing of the AELAW Database

In line with the main aim of this network, which is designing the future database of all the Palaeo-European fragmentary attested languages (excepting Latin, Greek and Phoenician; Figure 3.2), it is fundamental to create two censuses (for languages and for inscriptions, respectively) whose goal is not a critical edition of inscriptions, but only their quantification and identification.

This process has led to the recognition of approximately twenty languages and circa 20,000 inscriptions. Among the problems of linguistic identification, the most complex ones affect the Sabellic and Celtic branches (both problems will be faced in two conferences in 2018) and the indirect sources for the Balcanic languages, where only Thracian has been clearly identified. The best defined languages are Iberian, Celtiberian, Lusitanian, the “Southwestern (or Tartessian) language” in Hispania (plus Vasconic and Aquitanian, indirectly evidenced in both sides of the Pyrenees); Gaulish in France; and Lepontic in Northern Italy. In Italy and its islands: Elymian, Sicilian and Sikel are recognised in Sicily; Venetic, Messapic, Ligurian, Faliscan,

<sup>4</sup> [<http://aelaw.unizar.es/database/languages>].

Sabellian languages, Camunic, Raetic and Etruscan in the peninsula, being this last language the best represented of all, with more than 11,000 inscriptions.

### 3.4.2 Challenges Arising from the Digitalization of Palaeo-European Epigraphy and Solutions Addressed in AELAW

The particular nature of the AELAW database, developed as a census, has posed the following issues:

- The priority given to quantification, rather than to the content of inscriptions, has allowed resolution of a problem that was potentially unachievable in the census of inscriptions: the encoding of texts written in more than twenty writing systems. The language database does not pose any encoding problem.
- The creation of identifiers is particularly relevant in this stage. They will permit identification of the inscriptions whose fragments have been published in different moments, or the duplicated inscriptions (those whose fragments have been published as different inscriptions), the fake inscriptions, the inscriptions that are actually written in Latin, Greek or Phoenician. The collaboration with Trismegistos<sup>5</sup>, with which contact has already been established, will be fundamental in order to accomplish this task. The choice of the structure of the ID is at this moment a work-in-progress. Provisionally, the ID consists of the initial letter of the language of the text (e.g. Oscan=O, Venetic=V), allowing the user to clearly identify the epigraphic culture to which the text belongs, and a correlative number; but the team is currently assessing the possibility of assigning just a number as an ID of each inscription, just as Trismegistos does.
- The lexicographic solution of AELAW is relatively simple, compared to BDHesp. A working group specifically devoted to that task decided that the units of the languages database were languages, and the units of the inscriptions database were inscriptions, given that AELAW is mainly interested in the quantification of the data. Just as in BDHesp, both databases are linked to a third database containing the bibliographic references mentioned in the files.

We are firmly convinced that every progress made in the field of epigraphy, and Palaeo-European languages in particular, will be narrowly related to digital epigraphy, whose resources and potential must be fully exploited. This is what we believe after our experience with BDHesp.

---

5 [[www.trismegistos.org](http://www.trismegistos.org)]. See Chapter 15 in this volume.

## Bibliography

- Beltrán, F. & Jordán, C. (2017a). *Celtibérico. Lengua, escritura, epigrafía* (AELAW Booklet 1). Zaragoza: Prensas de la Universidad de Zaragoza.
- Beltrán, F. & Jordán, C. (2017b). *Celtiberian. Language, writing, epigraphy* (AELAW Booklet 1). Zaragoza: Prensas de la Universidad de Zaragoza.
- Estarán, M.J. & Beltrán, F. (2015). *Banco de Datos Hesperia de Lenguas Paleohispánicas (BDHESP). II. Numismática paleohispánica*. Vitoria: Libros UPV/EHU. Retrieved from [<https://web-argitalpena.adm.ehu.es/pdf/UHPDF151886.pdf>], 2017/12/12.
- Gorrochategui, J. & Vallejo, J.M. (forthcoming). The Parts of Hispania without epigraphy. In A.G. Sinner & J. Velaza (Eds.), *Palaeohispanic Languages and Epigraphies*. Oxford: OUP.
- Luján, E.R. (2005). Hesperia. The electronic corpus of Palaeohispanic inscriptions and linguistic records. *Review of the National Center for Digitization*, 6, 78–89.
- Orduña, E. & Luján, E.R. (2014). Implementing a database for the analysis of ancient inscriptions: the Hesperia electronic corpus of Palaeohispanic inscriptions. In T. Andrews & C. Macé (Eds.), *Methods and means for digital analysis of ancient and medieval texts and manuscripts* (Lectio. Studies in the transmission of texts and ideas 1). Louvain: Brepols.
- Orduña, E. & Luján, E.R. (forthcoming). Philology and technology in the Hesperia databank. *Journal of History, Literature, Science and Technology*. Retrieved from [<http://hesperia.ucm.es/Lujan-OrdunaAHLlist.pdf>], 2017/12/12.
- Orduña, E., Luján, E.R., & Estarán, M.J. (2009). El banco de datos Hesperia. In F. Beltrán, J. d'Encarnação, A. Guerra, C. Jordán, & B. Díaz (Eds.), *Acta Palaeohispanica X. Actas del X Coloquio sobre Lenguas y Culturas Paleohispánicas* (Palaeohispanica 9) (pp. 83–92). Zaragoza: Institución Fernando el Católico.
- Salomon, C. (2017a). *Raetic. Language, writing, epigraphy* (AELAW Booklet 2). Zaragoza: Prensas de la Universidad de Zaragoza.
- Salomon, C. (2017b). *Rético. Lengua, escritura, epigrafía* (AELAW Booklet 2). Zaragoza: Prensas de la Universidad de Zaragoza.
- Untermann, J. (1975–1997), *Monumenta Linguarum Hispanicarum* (Vols. 1–4). Wiesbaden: Reichert Verlag.
- Vallejo, J.M. (2016). *Banco de Datos Hesperia de Lenguas Paleohispánicas (BDHESP). III. Onomástica paleohispánica. I. Antroponimia y teonimia. 1. Testimonios epigráficos latinos, celtibéricos y lusitanos, y referencias literarias*. Vitoria: Libros UPV/EHU. Retrieved from [[https://web-argitalpena.adm.ehu.es/listaproductos.asp?IdProducts=UHPDF163064&titulo=Banco%20de%20Datos%20Hesperia%20de%20Lenguas%20Paleohisp%20nicas%20\(BDHESP\).%20III.%20Onom%20E1stica%20paleohisp%20nica.%20I.%20Antroponimia%20y%20teonimia.%201.%20Testimonios%20epigr%20E1ficos%20latinos,%20celtib%20E9ricos%20y%20lusitanos,%20y%20referencias%20literarias](https://web-argitalpena.adm.ehu.es/listaproductos.asp?IdProducts=UHPDF163064&titulo=Banco%20de%20Datos%20Hesperia%20de%20Lenguas%20Paleohisp%20nicas%20(BDHESP).%20III.%20Onom%20E1stica%20paleohisp%20nica.%20I.%20Antroponimia%20y%20teonimia.%201.%20Testimonios%20epigr%20E1ficos%20latinos,%20celtib%20E9ricos%20y%20lusitanos,%20y%20referencias%20literarias)], 2017/12/12.
- Velaza, J. & Moncunill, N. (2017a), *Iberian. Language, writing, epigraphy* (AELAW Booklet 3). Zaragoza: Prensas de la Universidad de Zaragoza.
- Velaza, J. & Moncunill, N. (2017b), *Ibérico. Lengua, escritura, epigrafía* (AELAW Booklet 3). Zaragoza: Prensas de la Universidad de Zaragoza.
- Wodtko, D. (2000), *Monumenta Linguarum Hispanicarum* (Vol. V.1). Wiesbaden: Reichert Verlag.
- Wodtko, D. (2017a). *Lusitanian. Language, writing, epigraphy* (AELAW Booklet 4). Zaragoza: Prensas de la Universidad de Zaragoza.
- Wodtko, D. (2017b). *Lusitano. Lengua, escritura, epigrafía* (AELAW Booklet 4). Zaragoza: Prensas de la Universidad de Zaragoza.

Francesco Di Filippo

## 4 Sinleqiunnini: Designing an Annotated Text Collection for Logo-Syllabic Writing Systems

**Abstract:** Sophisticated writing systems, such as Cuneiform and Linear B, pose tremendous challenges for the development of digital corpora of annotated textual documents. The fact that both of them do not clearly represent the spoken form of the underlying languages, as well as the multi-level character of their logo-syllabic writing systems, has required the setting up of an *ad hoc* solution for complex data handling, aimed at capturing all of their features. While the usual approach of adapting a mark-up language would have been possible at least in principle, Sinleqiunnini relies on a different formal model, having been conceived from its beginning as a database driven framework. Such a solution was demonstrated to be more efficient than mark-up languages in representing parallel, overlapping hierarchies, while it also simplified prototyping of a set of complex queries to exploit the different information levels of these texts. Finally, it provided a more functional instrument to perform multi-user/multi-level annotation.

**Keywords:** Cuneiform, Linear B, relational model, XML schema, mark-up languages

### 4.1 The Project

Sinleqiunnini aims to be a software framework for the management of digital repositories of epigraphical sources, primarily concerned with logo-syllabic writing systems from the eastern Mediterranean basin, and their dissemination through the World Wide Web.<sup>1</sup>

The early stage of the project, which originated in 2006 under the supervision of C. Zaccagnini at the University of Naples “L’Orientale”, focused on the setting up of a digital repository to store, visualize and query a textual database of cuneiform tablets from Emar (Syria), which was encoded as pure text files in the late ’80s.<sup>2</sup> During these early developments, the project’s primary objective was the setting up of a digital representation of texts to fit in the current transliteration methodology in order to

---

1 [[www.pankus.com](http://www.pankus.com)].

2 [<http://virgo.unive.it/emaronline/cgi-bin/index.cgi>].

---

Francesco Di Filippo, Consiglio Nazionale delle Ricerche, Roma



© 2018 Francesco Di Filippo

This work is licensed under the Creative Commons Attribution 4.0 International license (CC BY 4.0)

render the digital edition of texts virtually indistinguishable from their original printed layout. Since its beginning, it has been conceived to be Unicode aware and it was also one of the first projects dealing with cuneiform sources that bypassed the workaround of reproducing online editions through special, often unreadable, complex ASCII pseudo-encoding. This choice forced us towards high-level technical solutions that could efficiently manage variable-width length character encoding such as UTF-8. In 2006 the choices were quite restricted, so that we confidently relied on MySQL for data persistent storage and Perl as scripting language. At that time, the data model was a simple collection of occurrences of all lexical entities having been earlier tokenized from pure text files. Yet, even at this early stage of development, the software was quite efficient and responsive. Besides the capability of representing texts as in their printed layout, this first project already allowed search by string matching and regular expressions, in order to extract meaningful patterns by context, syntagmata and co-occurrences, and to produce glossaries of the digital collection.

Over time, Sinleqiunnini developed intermittently until it faced new, specific challenges arising from two very different textual collections. In 2011 the framework was employed to develop LiBER (Linear B Electronic Resources), a CNR project realized in collaboration with M. Del Frio, which aimed at producing a digital edition and a query tool for the Linear B documents (Del Frio & Di Filippo, 2014).<sup>3</sup> During this phase, Sinleqiunnini's architecture expanded with additional modules that were introduced in order to address issues concerning the spatial distribution of epigraphic phenomena, thus modifying the earlier data model architecture by enriching the system through Web-GIS capabilities.

In 2015, the data model architecture underwent a radical restyling. From 2008, the project was already in use for the management of the digital edition of the entire corpus of cuneiform texts belonging to the Ebla royal archives, EbDA (Ebla Digital Archives), a project of University of Venice "Ca' Foscari" in collaboration with L. Milano and M. Maiocchi.<sup>4</sup> During this last project development, we benefited from the extraordinary contribution of R. Orsini, who helped us develop a brand new relational scheme, the one in use today in Sinleqiunnini. This new implementation, which constitutes the object of the present article, has greatly enhanced database performances, while providing more effective querying and data-mining perspectives. More significantly, it also allowed the increase of the granularity of the database model, giving access to the management of the collection at its very basic unit level (i.e. cuneiform signs), and contributed to the design of a more consistent solution for multi-level/multi-user annotations (Di Filippo et al., 2018).

---

<sup>3</sup> [<http://liber.isma.cnr.it>].

<sup>4</sup> [<http://ebda.cnr.it/>].

## 4.2 Collection Design: Mark-Up Languages Versus Database Model

In a seminal article of 1990, with the purpose of designing a standard for encoding machine-readable documents, DeRose et al. (1990) boldly introduced the notion of “content object” as a logical structure of a document, “having to do with meaning and communicative intention”. In the same contribution, they defined the document itself – i.e. its digital form – as a representation of an “ordered hierarchy of content objects” (the so-called OHCO model). In this view, a document is essentially the product of the juxtaposition of a series of nesting objects such as chapters, paragraphs, words, and so on, each of them containing elements of lower order. In the early ’90s, this model was by far the simplest and most functional way to create, modify and format texts. Digital documents were represented in this way to support browsing, text mining procedures, and other sorts of special processing, and they were much more easily shared among different applications and platforms. It is not by chance, then, that this “ordered hierarchy of content objects” proved to be an effective premise in pushing the use of descriptive mark-up languages to represent digital documents. More specifically, it provided the most advantageous theoretical framework for projects involved in humanities computing, such as the Text Encoding Initiative (TEI).

Over time, however, some of the authors of the original thesis have identified a basic flaw in the apparent simplicity of the OHCO model. A textual document is indeed more often the result of several logical structures, a series of hierarchies that can also be reasonably considered “logical” (Renear, Mylonas, & Durand, 1993). By addressing the problem from different analytical perspectives, it soon emerged that a text may in fact have concurrent, overlapping hierarchies, and that this kind of textual source cannot be easily represented by a tree-shaped data structure. “Non-nesting information poses fundamental problems for any XML-based encoding scheme, and it must be stated at the outset that no current solution combines all the desirable attributes of formal simplicity, capacity to represent all occurring or imaginable kinds of structures, suitability for formal or mechanical validation. The representation of non-hierarchical information is thus necessarily a matter of trade-offs among various sets of advantages and disadvantages”.<sup>5</sup>

Another important drawback in the adoption of a descriptive mark-up language for the architecture of a large repository of texts is deeply rooted in the metadata management. Any kind of information not directly belonging to any given hierarchy – i.e. extra-textual information such as metadata – can be tied only to the same structure of the text and must be expressed as a string of the mark-up language. This quite impractical restriction often pushes back-end developers towards the use of alternative data containers for persistent metadata storage. It is not uncommon to

---

5 [<http://www.tei-c.org/release/doc/tei-p5-doc/it/html/NH.html>].

meet mixed solutions indeed, solutions that pair mark-up languages for texts with relational databases for metadata. Such mixed workarounds are in use to such an extent that, as an apparent contradiction in terms, a giant of relational database management such as PostgreSQL since long (version 8.3) has been forced to introduce ways of storing loosely structured data like XML.<sup>6</sup>

Having discussed two of the main pitfalls in adopting a descriptive mark-up language in encoding textual sources, it is important to address more strictly some of the issues concerning the architecture of our project in relation to the peculiar type of sources it deals with.<sup>7</sup>

Consider, for instance, the case of a clay tablet, be it drafted through the archaic cuneiform of Ebla, or through the Linear B writing system. At least two concurrent, overlapping hierarchies may represent the structure of the document. There exists, in fact, a physical structure such as tablet > lines > words or, as in the case of the administrative documents from Ebla, a more complex structure such as tablet > columns > boxes > lines > words (see *infra*). These hierarchies overlap a further structure, that is the logical representation of the document such as text > paragraphs > words. This document has a title (e.g. MY Ue 652+656 or ARET 1 1), and may be enriched with information about the archaeological context of each of the fragments that constitute the document. This level of information (i.e. metadata) – although may be represented as a nesting structure as well (e.g. site > building > room > findspot) – does not belong to any of the hierarchies of the document and is far better represented by a relational model, whose ultimate goal is preserving data consistency and diminishing redundancies. This document may eventually be annotated, both with grammatical categories in the shape of tree-structured data and with commentaries made by different scholars, which over-time have given different interpretations and readings to some of the text's passages.

Parallel to this structure of our abstract sample text – a structure quite common of any digital collection of historical sources – further levels of information arise by addressing the peculiar nature of logo-syllabic writing systems. At the most general level, the documents of the digital collections considered here, record information by means of a rather large set of glyphs, usually ranging from a couple of hundred items up to a couple of thousand, depending on time, region, and text corpora. Within these writing systems, signs may be defined by functional classes that more or less

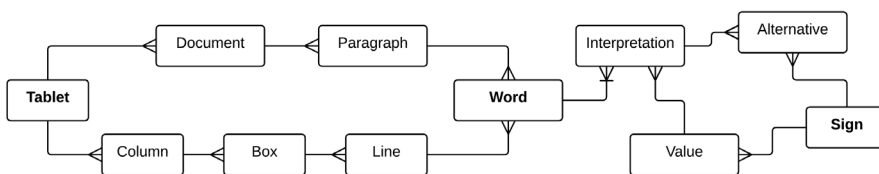
---

<sup>6</sup> [<https://www.postgresql.org/docs/8.3/static/datatype-xml.html>].

<sup>7</sup> During the past few years, we witnessed the emergence of a considerable number of projects involved in digital editions of cuneiform corpora (Charpin, 2014). However, notwithstanding relevant drawbacks in the use of XML based model even for modern alphabetic scripts, most of them relies precisely on the usual approach of adapting a mark-up language. As regards the Linear B writing system, instead, the problem concerning the management of annotated textual collections has been taken into consideration from a different perspective. Hitherto, the only two projects focusing on these sources rely on a database-driven approach (Del Freato & Di Filippo, 2014; Aurora, 2015).



correspond to the base classes of syllabograms, logograms and determinatives. In the case of Linear B, signs are usually more specialized than in the cuneiform writing system (Del Freato, 2016a). In the latter, a given glyph is often associated with more than one function, whose value sometimes can be revealed only by the context (Reiner, 1966). For instance, the “earth” sign, besides being used as a determinative for geographical names (usually rendered by superscript characters), may also stand for the word “earth” or “place” (respectively *eršetu* and *ašru* in Akkadian), or for the syllabic value /ki/ of the personal name *a-bar-ki*. It is of note that the latter sign sequence deeply relies on its context: it can stand for either a geographical name (*a-bar<sup>ki</sup>*, e.g. ARET 15 32) or a personal name (*a-bar-ki*, ARET 8 522, a name indicating a professional qualification). Furthermore, a given sign may be associated with more than one logographic value and also with more than one syllabic reading. For instance, the KA sign may be read ka “mouth”, zu<sub>2</sub> “tooth”, inim “word”, etc., whereas the sign GA may be used to represent the syllables /ga/, /qa/, /ğa/, according to the so-called polyphony principle. Conversely, two or more different signs may end up having the same reading (homophony principle); in this case they are conventionally distinguished in modern transliterations by a lowercase numerical index (or accents). All this, at the more abstract level, entails the possibility of an unpredictable number of graphic variants of the same and unique word. From the perspective of the data model design, this also entails the necessity of enriching the structure of digital text with at least two further concurrent, overlapping hierarchies: the one bearing the actual reading of the text (i.e. its interpretation), the other recording un-interpreted graphemic sequences of conventional signs’ names (Figure 4.1). In addition, scribal mistakes, signs added by modern editor, palimpsests, or erasures, often entail the addition of further nested structures for the management of the document’s minimal units.



**Figure 4.1:** Synoptic scheme of parallel hierarchies

Another peculiarity of the cuneiform writing system is that graphemes, i.e. distinct minimal units within the sign corpus, may be arranged in a number of different ways, by: inclusion (partial or total), juxtaposition, ligature, crossing, and so on. For instance, the sign KU<sub>2</sub>, which is used to express the verb “to eat”, is composed by two graphemes: the sign for “food” (originally a pictographic representation of a vessel

for rations) either within the sign for mouth, or in close proximity to it. However, in the latter case an interpretation in terms of a different compound logogram, namely *inim gar* “to make a legal claim” (lit. “to place/put a word”), is also possible. This is an extreme case, but uncertainties in the interpretation of the documents may suggest leaving the readings of some sign or sign sequence open.

While the usual approach of adapting a mark-up language, like SGML and XML, perhaps using an EpiDoc-based encoding, could be possible at least in principle (DeRose, 2004; Smith, 2007; Witt & Metzger, 2010),<sup>8</sup> such a document would be very complex to create, manage and use. Even with sophisticated tools, an approach based on a descriptive mark-up language would require the development of *ad hoc* solutions, at the expense of greatly increasing the complexity of the representation and making even more difficult both the access to the textual collection and the processes of information retrieval (Iacob & Dekhtyar, 2005).

Some of the issues posed by logo-syllabic writing systems, to my knowledge, cannot be addressed by an architecture based on a descriptive mark-up language. The main limit in adopting, for instance, an XML scheme for a digital repository of transliterated primary sources from the ancient Near East and Aegean logo-syllabic scripts, in fact, concerns the impractical representation of complex, structured annotations. In XML, because attributes can only be represented by plain text and are directly bound to single elements of the hierarchy, it is virtually impossible to annotate non-contiguous portions of text. Moreover, annotations cannot overlap, nor is it possible to annotate concurrent hierarchies within the same and unique instance. These drawbacks in the annotation procedures of mark-up encoding systems, conversely, represent an essential prerequisite in the design of a digital collection planned for a variety of studies on the development of some of the earliest writing system in the history of mankind. Two examples will probably better clarify these very specific needs.

The layout of cuneiform and Linear B primary sources is very different from the literary documents for which mark-up systems were originally conceived. Linear B tablets do not pose many problems in this regard (Del Frego, 2016b). The writing system, moving from left to right by superimposed lines, more or less parallels a modern layout. The second millennium cuneiform system from Emar, at a great extent, arranges information in lines of even size and is also comparable to the modern stream of text, but some important exceptions may occur. For instance, on a consistent number of legal tablets of the Syro-Hittite scribal school (Seminara, 1998) the document ends with a series of seal impressions with Luwian hieroglyphic inscriptions, to which cuneiform legends bearing name and patronymic of the seal’s owner are associated. The shape of these constructs is synoptically rendered as follows:

---

<sup>8</sup> In the scope of cuneiform studies, it is worth to cite the ORACC project [<http://oracc.museum.upenn.edu/>].

.1	PN1 [seal impression]	PN3 [seal impression]
.2	son of PN2	son of PN4

As it is possible to observe, the natural stream of the text does not conform to the logical structure of the document one may want to annotate. Indeed, the primary level of information is not the relationship between the PN1 and PN3, which in fact lay one after the other, both on the original cuneiform tablet and in its digital reproduction. The main researcher's concern, potentially in order to analyse prosopographic ties, is actually the relationship between father and son, whose names are separated by the seal impression in the natural stream of the cuneiform tablet. As a result, in these cases, if one would like to save the original layout of the document, it would be necessary to annotate non-contiguous portions of text by putting together hundreds of these occurrences, a procedure that – even if possible in XML – would compromise human readability of the output and would require special tools to be fruitfully handled.

The second example of problems of practical annotation concerns a frequent feature occurring in the administrative archives of Ebla. The cuneiform writing evidenced in this site of inner Syria of the third millennium BCE shows many archaic features, especially in the layout of the administrative tablets, which form the bulk of the archive. Text is usually arranged in columns – to be read from top to bottom, from left to right – each containing several boxes, which in turn are inscribed with lines of uneven size. Each box usually contains a semantic unit such as, for instance, a number plus the noun it refers to, a verbal form, a preposition, and so on. In those documents, some items, which are very frequently evidenced (such as  $ib_2$  “belt” that occurs more than ten thousand times), appear into a variety of “crystallized” graphemic sequences in which the order of the elements does not conform to any linguistic scheme. In this respect, the sequence  $ib_2$ -III-dar- $sa_6^{tug_2}$ , often transliterated with hyphens between each word unit (despite some variants that may occur, due to different editors' preferences), can be interpreted as follow:

- $ib_2$  > the main lexeme for “belt”;
- III > a numeric attribute, probably denoting its length;
- dar > a qualifying adjective referring to the main lexeme, to be read “colored”;
- $sa_6$  > a further adjective, meaning “of good quality”;
- $tug_2$  > the determinative for this class of objects, that is “textiles”.

There is more than one apparent incongruence in the way such a linguistic unit is represented. First, one would expect the three adjectives denoting the character of the “belt” (III, dar,  $sa_6$ ) to be separated one from the others and from the qualified lexeme ( $ib_2$ ) by means of a white space. Second, the determinative should immediately follow the lexeme and should not be placed at the end of the sequence: it determines the nature of the “belt” as a textile and not, as in this case, the last adjective in the chain

of morphograms. In short, this is a further example of the necessity of a flexible instrument to manage annotations of non-contiguous lexical units.

In addition, it should be stressed that exactly the same compound semantic unit is more often differently rendered, most frequently as a modern reader would expect (e.g.  $ib_2-II^{tus}$ :  $sa_6$  dar, ARET 1 1, r.4,7), but also by means of more convoluted sequences, such as: 4  $ib_2-IV$   $sa_6$  dar 5  $ib_2-III$  dar  $tug_2$  (e.g. ARET 1 1, r.3,6). In this case, we have probably the clearest example of the difficulties in projecting these compound semantic units into linear patterns – which nevertheless is the common praxis for printed layout. In the latter example, the ancient scribe wrote the determinative for textiles (i.e.  $tug_2$ ) at the end of the line (i.e. of the box), clearly with the intent of qualifying different semantic units of the same type with one determinative only. The typographic rendering of the sequence, the only viable option to preserve the integrity of the original document, however, has the counter-effect of generating a sort of linguistic ambiguity. Given the multi-level nature of the cuneiform writing system, an isolated sign  $tug_2$  at the end of the line could even be considered as an independent linguistic unit. It would not qualify the nature of the preceding “belts” as textiles, but it would be considered as an independent word meaning “dress”, thus distorting the sense of the whole sentence.

In order to better preserve the original textual layout and, at the same time, to safeguard the essential underlying level of information, it is then necessary to conceive a conceptual scheme that would allow annotations of non-contiguous lexical units and, as in the above-mentioned case, a system in which different instances can overlap without conflicting. In this regard, Sinleqiunnini allows annotations of any type, be they strings, structured values or references to other sections of the document. Being able to reference multiple textual objects as a single entity and, above all, to work with overlapping textual objects, it allows the user to annotate even arbitrary portions of the document. Going back to the above-mentioned sequence, the issue posed by this semantic pattern can be easily solved by referencing the determinative  $tug_2$  twice as an instance of both the two preceding textual objects:

---

- original sequence:	4 $ib_2-IV$ $sa_6$ dar 5 $ib_2-III$ dar $tug_2$
- underlying annotation:	4 $ib_2-IV^{tus}$ : $sa_6$ dar 5 $ib_2-III^{tus}$ : dar

---

Finally, a further topic has been considered and technically solved by Sinleqiunnini’s architecture. Logo-syllabic texts, as may be inferred by the writing system outlined above, are characterized by a substantial level of uncertainty and variation. Some text interpretations either rest on different scholars’ readings, sometimes conflicting, or are still unavailable. Thus, the building of an integrated digital collection must envisage a multi-user, multi-level annotation system, in order to keep track of this set of overlapping interpretations. Yet, for the reasons discussed above, this system

cannot be easily handled by any of the mark-up languages commonly in use for the representation of large textual collections.

### 4.3 Sinleqiunnini Data Container

The project, since its latest development, greatly benefited by having been ported to Python as scripting language, Flask as framework web, and PostgreSQL as the relational data management system. In addition, in order to facilitate the conversion between incompatible type systems, the data container structure has been re-organized through an object-relational mapping system (ORM), namely SQLAlchemy (Myers & Copeland, 2015). This allows the project to interact with regular Python objects instead of working with database entities such as tables, documents, or Structured Query Language (SQL), yet it allows mixing use of the ORM with the SQL to satisfy very specific issues.

This substantial rewriting of the project source code has been the occasion to address, in a more formal manner, problems of data persistency by formalizing a new conceptual schema. This new schema has been deeply influenced by the high level of formalism of the Manuzio project (Maurizio & Orsini, 2010a), from which Sinleqiunnini differs in the structure of implementation (Maurizio & Orsini, 2010b).

Both the influence of Manuzio and the adoption of an ORM system have greatly contributed to the rethink of the nature of textual collections. Quite surprisingly, the above-mentioned OHCO model, which considers text as “ordered hierarchies of content objects”, still proved to be the most serviceable theoretical framework for designing multi-layer textual documents. However, the many problems of adopting standard mark-up language solutions and, above all, the multi-level structure of non-alphabetic textual sources, led us to design an *ad hoc* solution for the management of these hierarchies of “content objects”.

The core concept of our project’s architecture is that a text can be represented as a set of hierarchies of either *textual* or *association objects*.

*Textual object* is an abstract representation of the different logic structures that contribute at defining a text as such; it has a logical meaning such as line, paragraph, word, sign, and so on. In other terms, a textual object is the sum of the portion of text with its structural (i.e. object’s properties) and behavioural aspects inherited by ORM logic. Those aspects are of great help in maintaining data consistency. Textual object behaviour, in other terms, is a set of local procedures (i.e. methods), which help define computed properties, as well as perform operations on the represented portion of text. For instance, any time some text value is sent to the database, pertinent methods can check the validity of the information by validating it against a set of dictionaries (e.g. a syllabary) previously defined for the collection.

An *association object* has a slightly different nature, as well as a higher degree of abstraction: because textual objects cannot contain duplicates, association objects

are intended to keep track of positional and contextual information of textual objects. For instance, in our document collections a textual object “Tablet” is intended to represent an instance of a physical document, e.g. MY Au 102, alongside all its attributes and references to lower-order logical structures. Obviously, being the highest order item in the hierarchy, there is only one instance of this type in each collection. Lower-order objects, anyway, need to be repeated as many times as the actual occurrences of these objects. The Mycenaean tablet MY Au 102 has references to 15 instances of the textual object “Line”, to 35 instances of the textual object “Word”, and so on. In other terms, a text may consist of many lines and a line may consist of many words: in relational model jargon, this is a typical example of a one to many relationship, which in turn is a perfect representation of a hierarchy by nested structures. However, it is also important to point out that in a textual document the same words may recur, sometimes with a very high frequency. In the case of our sample text, the Linear B logogram for “man” (by convention rendered by Latin word “VIR”) appears 9 times in MY Au 102, thus representing more or less 26% of all the words of the above document. Do we really need to separately record each instance of the same word for “man”?

The relational model on which Sinleqiunnini rests allows a more convenient way to keep track of such information. The two textual objects, “Line” and “Word”, were conceived to reference each other through an association object (i.e. “Occurrence”), which in turn is intended to permanently store the position of each of the unique occurrences of lines and words. In other terms, the logogram VIR exists only as a unique instance of the object “Word”, but its position within the lines of the tablet is duly recorded as a numeric index by the Occurrence association object. The latter, moreover, is conceived to collect all the contextual attributes of its referenced textual object, attributes that may characterize the nature of the underlying finite sequence of characters (i.e. string) at a given position in the document.

The same is true for very frequent terms, as in the case of the Eblaite word  $'a_3$ -*da-um* (i.e. some kind of cape). To better illustrate this example, it is necessary to introduce a further characteristic of our data model. Sinleqiunnini’s hierarchy of words rests on the difference among epigraphic notations (“Notation”), words (“Word”), and lexical entries (“Lemma”). The first textual object is intended to record all the possible, different forms in which a given term may occur, accounting for those signs not belonging to the original text and introduced by the critical edition to preserve a level of information concerning the physical state of the source (e.g. the square brackets for fractures). Of course, these characters are necessary to keep the digital representation of the document as close as possible to its printed layout, but they may hamper searching operations and comparison between terms. This is the main reason for the introduction of the abstract textual object Word. This collects unique instances of words from which all the editorial markers have been removed: thus, the two notations  $'a_3$ -*da-[um]* and  $'[a_3-d]a-u[m]$  refer to the same word  $'a_3$ -*da-um*. A third textual object, Lemma, is intended to archive headwords of the inflected terms (i.e. canonical form or dictionary form), thus providing the system with a higher-level

clustering property. Turning back to the above example of association object, thus, Sinleqiunnini stores the very frequent Eblaite term  $'a_3$ -*da-um* as instances of three different nesting textual objects. At a higher level, there exists only one instance of the Lemma  $'\grave{a}$ -*da-um*, eventually enriched with a set of attributes for its translations into modern languages. Then, there are multiple instances of the Word textual object such as  $'a_3$ -*da-um-I* or  $'a_3$ -*da-um-II* (by praxis, the base term and its specific numeric attribute are always treated as a compound element); finally, there are several Notation instances, as many as the single occurrences of its epigraphic notation variants. For instance, for the term's transliteration  $'a_3$ -*da-um*<sup>tug2</sup>, that is a single lemma instance, more than one hundred different epigraphic words exist, which in turn refer to more than five hundred notations of the Eblaite word for “cape” in the collection of texts currently available.

Such a level of simplification has a significant impact on textual collection management. Any time the philological and epigraphical research provides a new reading for a given graphemic sequence – and this happens quite frequently in cuneiform studies – it is sufficient to update a unique instance of the object at the word level in order to make this change propagate by cascading effect on the textual collection as a whole. Moreover, it has relevant consequences in terms of searching and pattern matching procedures: Sinleqiunnini's search engine has to process only one item for each user's query. This resulting object, being characterized by the principle of inheritance of the object-oriented language, however, is intrinsically enriched by all its relationships with referenced objects, as well as by all pertinent positional and contextual information.

A last, concluding remark focuses on prototyping a multi-user and multi-level architecture to provide the system with cooperative annotations capabilities. In Sinleqiunnini, given that the structure of the textual object is capable of referring to arbitrary portions of the underlying text, annotations can be attributes of any type, be they strings, structured values or references to other textual objects. From this perspective, annotations are logical structures that can also encompass non-contiguous sets of lexical entities and, unlike mark-up language approaches, can overlap without the risk of conflicting. In addition, the relational database architecture provides researchers with the most efficient background for the management of multi-level sets of annotations.

The fact that text readings often rest on conflicting interpretations of different scholars poses remarkable challenges as concerns the number and dimension of annotations to be collected for each textual entity. Consider, for instance, the following excerpt from the Emarite tablet RAE 202:

ll. 13-14:  $u_3$  *a-nu-ma* *ṭup-pa* [š]a E<sub>2</sub> <sup>4</sup>IM *ma-ri* // <sup>t</sup>*tar-ši*<sub>2</sub>-*pi*<sub>2</sub> *il-t[a-qu]* (Arnaud, 1986)  
(the tablet of the temple of god Ba'al, the sons of Turšipu have taken).

This cuneiform tablet has been the object of several studies. Over time, very different readings have been proposed for these two lines, deeply conditioning historical research. From our perspective, these concurrent levels of information entail the necessity of a flexible annotation tool, not least because it is not yet possible to select a preferential interpretation for this text:

- 1)  $u_3$  *a-nu-ma tu<sup>p</sup>-pa ša E<sub>2</sub><sup>md</sup>IM-ma-<lik ma>-ri // <sup>f</sup>tar-ši<sub>2</sub>-pi<sub>2</sub> il-t[a-qi<sub>3</sub>]* (Durand & Marti, 2003)<sup>9</sup>  
 2)  $u_3$  *a-nu-ma tu<sup>p</sup>-pa ša E<sub>2</sub><sup>cm>d</sup>IM-ma-lik! // <sup>f</sup>tar-ši<sub>2</sub>-pi<sub>2</sub> il-t[a-qu]* (Cohen, 2009)  
 3)  $u_3$  *a-nu-ma tu<sup>p</sup>-pa ša <sup>˘</sup>E<sub>2</sub><sup>md</sup>IM-ba-ri // <sup>f</sup>tar-ši<sub>2</sub>-pi<sub>2</sub> il-t[a-qi<sub>3</sub>]* (Yamada, 2013)

---

1) the tablet concerning the house of <b>Ba'al-malik, the son of Turšipu has taken.</b>	2) the tablet concerning the house of <b>Ba'al-malik, which Turšipu took, ...</b>	3) the tablet concerning the house of <b>Ba'al-baru, Turšipu has taken.</b>
---	---	---

---

These four readings (that one of the tablet's first editor and the three new interpretations) are, except one, the result of the juxtaposition of the same number of tokens. In the interpretation no. 1, indeed, the assumed omission of two signs led the authors to split the personal name into two tokens, thus altering the paragraph length. As a consequence, when single word readings are different, it is impossible to simply collect these variants as attributes of the base instance of a word-level textual object. In Sinleqiunnini, instead, all those interpretations are intended as discrete logical units and these units are referenced to a common textual object type "Paragraph". As a consequence, alongside the basic reading of the document (eventually the one proposed by the original editor), there exist at least three parallel interpretations that potentially can be selected in the web-based user interface. At the same time, different instances of word-level objects, down to the collection of the minimal unit (i.e. cuneiform sign), are also referenced to the different interpretations, so that it is possible to perform searches even for these parallel discrete logical units. The resulting output then will specify the provenance of a given lexical entity and, eventually, if this word is part of an alternative reading proposal.

Finally, via bibliographic references, each new reading proposal is intrinsically tied to different scholar's authorities, which may also help end-users select pertinent interpretations for highly controversial text passages.

---

<sup>9</sup> The <> markers stand for a modern insertion of cuneiform signs.



## 4.4 Conclusions

The complexity of the logo-syllabic writing systems offers stimulating challenges to specialists in philology, information technology, and digital humanities alike. As digital humanities positively impacts on all fields involved in the study of the past, it becomes increasingly clear that traditional research methodologies must be matched by state-of-the-art research tools. The development of innovative instruments is, however, a slow and expensive process. It requires close cooperation of experts in diverse fields, which in turn rests on the creation of a common, cross-domain language in order to facilitate this interplay. In order to minimize these drawbacks, it is important for philologists - and more generally, for researchers of the human past - to develop hybrid expertise, which would greatly help this dialogue with the information technology world. This would also greatly benefit their potential as scholars, as basic knowledge of data management and scripting languages may open up lines of research that would otherwise remain unexpressed. This is due not only to the greater paucity of financial resources, but predominantly because of a lack of vision of this complex system as a whole. It is the fertile interplay of these newly established scholarly domains that make significant advancements in the understanding of our history possible.

It is exactly with this spirit in mind that the Sinliqiunnini project has developed, although intermittently, during these last ten years.

The project has defined a data model capable of representing the complexity of logo-syllabic writing systems by storing more information (and in a more useful way), compared to previous digital corpora of such a genre. Likewise, through this system, more sophisticated queries and analysis are possible due to the fact that data can be re-aggregated, on a case-by-case basis, through specific “views”, which may reflect more strictly the specific needs of a given line of research. This, of course, relies on the fact that our approach rests on database technology, and the fact that our data model is not directly bound to any given textual hierarchy. Conversely, in our system each hierarchy, each ordered juxtaposition of logical structures, has the reasonable claim to be the fundamental digital representation of the document. There is no more need to “simply to pick a single hierarchy as the ‘real’ document hierarchy, and flatten all other hierarchies” (Renear, Mylonas, & Durand, 1993). The numbers of these equally important nested structures can grow over time without any significant impact on previously created querying tools or on the coherence of the collection as a whole. Since the structure of Sinliqiunnini is thought as a modular sequence of Textual Objects managed by a relational database, and not as a mere text file, each structure of the document is separated from the others and new Textual Objects can be added. New, logical structures can enrich the digital collection, also new structures that may not have been foreseen during the design phase of the digital repository.

This is the reason why the system (although in this regard it is still in its early stages of development) has been able to introduce an innovative annotation system, capable of bypassing intrinsic limits of the XML schemes, which will support the

collaborative work of scholars, enhancing the information contained in the database via annotations.

Finally, our data model supports a set of sophisticated data extraction and analysis operations:

- advanced queries based on regular expressions, matching any of the following: part of a word, whole word, word starting with, word ending with; user defined input string formatted according to PostgreSQL regular expressions syntax;
- Full Text queries on English translations – based on stemming (ex: a query for “goes” returns “to go” as well);
- queries on ancient lexical roots associated with the individual words, based on the Textual Object Lemmas.
- queries for syntagmatic units: match one or more input strings within a user-defined word range – e.g.: match the word for “house” (E<sub>2</sub>) only when it is followed by the word for “king” (EN); match the word for “king” only when it is mentioned together with the word for “queen” within an interval of two words (e.g.: “king and queen”);
- co-occurrences: match texts containing an array of words – e.g.: a list of city names. This comes with a further option, namely an exclusion list – e.g.: match all texts containing both Ebla and Kakmium, but not Mari;
- queries for sign names: given an input reading, match all possible values attached to the corresponding sign. If two or more readings are passed as input, the query returns all words containing the corresponding input signs attached to them, regardless of their actual readings. Depending on user preference, the input string matches either two or more consecutive signs, or signs within a user-defined range.

During the past few years, we witnessed the emergence of a considerable number of projects involved in digital editions of cuneiform corpora (i.e. Charpin, 2014). Paradoxically, the tremendous amount of work has been perceived as something considerably different from traditional printed editions. Part of the issue is related to the actual evaluation system for the research products, which in EU countries at least is not yet capable of adequately evaluating the impact of state-of-the-art online digital tools, which are *per se* research products. Another part of the issue may be related to the fact that current online projects show a very high degree of variability. Most of them opted for proprietary conventions for the digital representation of their contents, either adapting an existing mark-up language or setting up an original one. Despite some relevant results, however, this process has hampered one of the prerequisites of the scientific research, which is the possibility of sharing information and data among scholars. We believe that it is time for modern philologists to consider the significant need for adoption of a shared digital grammar (encoding, data model, platform, query tools), specifically conceived for the management of the complexities of the logo-syllabic textual sources. We hope our project may serve as a starting point

in the definition of such grammar, to be further refined in order to assess the specific points of interest of the individual projects.

## Bibliography

- Arnaud, D. (1986). *Emar VI.3. Textes sumériens et accadiens*. Paris: ERC.
- Aurora, F. (2015). DĀMOS (Database of Mycenaean at Oslo). Annotating a Fragmentarily Attested Language. *Procedia - Social and Behavioral Sciences*, 198(1877), 21–31.
- Charpin, D. (2014). Ressources assyriologiques sur Internet. *Bibliotheca Orientalis*, 71, 331–357.
- Cohen, Y. (2009). *The Scribes and Scholars of Emar: Ancient Scribal Education in a Late Bronze Age City*. Winona Lake, Indiana: Eisenbrauns.
- Del Freo, M. (2016a). La scrittura lineare B. In M. Del Freo & M. Perna (Eds.), *Manuale di epigrafia micenea. Vol. I* (pp. 123–166). Padova: Webster.
- Del Freo, M. (2016b). Classificazione dei documenti e regole di trascrizione. In M. Del Freo & M. Perna (Eds.), *Manuale di epigrafia micenea. Vol. II* (pp. 247–256). Padova: Webster.
- Del Freo, M. & Di Filippo, F. (2014). LiBER: un progetto di digitalizzazione dei testi in scrittura Lineare B. *Archeologia e Calcolatori*, 25, 33–50.
- DeRose, S.J. (2004). Markup Overlap: A Review and a Horse. In *Proceedings of Extreme Markup Languages, Montréal*. Retrieved from [http://conferences.idealliance.org/extreme/html/2004/DeRose01/EML2004DeRose01.html], 2017/11/20.
- DeRose, S.J., Durand, D.G., Mylonas, E., & Renear, A.H. (1990). What is text, really? *Journal of Computing in Higher Education*, 1(2), 3–26.
- Di Filippo, F., Maiocchi, M., Milano, L., & Orsini, R. (2018, in press). The “Ebla Digital Archives” Project: How to Deal with Methodological and Operational Issues in the Development of Cuneiform Texts Repositories. *Archeologia e Calcolatori*, 29, 117–142.
- Durand, J.-M. & Marti, L. (2003). Chroniques du Moyen-Euphrate 2. Relecture de documents d’Ekalte, Émar et Tuttul. *Revue d’assyriologie et d’archéologie orientale*, 97, 141–180.
- Iacob, I. & Dekhtyar, A. (2005). Towards a Query Language for Multihierarchical XML: Revisiting XPath. In *Proceedings of the 8th International Workshop on the Web and Databases (WebDB 2005)* (pp. 49–54). Baltimore, Maryland: Citeseer.
- Maurizio, M. & Orsini, R. (2010a). Manuzio: a model for digital annotated text and its query/programming language. In M. Lalmas, J. Jose, A. Rauber, F. Sebastiani, & I. Frommholz (Eds.), *Proceeding ECDL’10 Proceedings of the 14th European conference on Research and advanced technology for digital libraries* (pp. 478–481). Berlin: Springer.
- Maurizio, M. & Orsini, R. (2010b). A Model and a Language for Large Textual Databases. In S. Bergamaschi, S. Lodi, R. Martoglia, & C. Sartori (Eds.), *Proceedings of the Eighteenth Italian Symposium on Advanced Database Systems, SEBD 2010* (pp. 254–265). Bologna: Esculapio editore.
- Myers, J. & Copeland, R. (2015). *Essential SQLAlchemy* (2nd ed.). Sebastopol, CA: O’Reilly Media.
- Reiner, E. (1966). *A Linguistic Analysis of Akkadian*. London - The Hague - Paris: Mouton & Co.
- Renear, A.H., Mylonas, E., & Durand, D. (1993). Refining our Notion of What Text Really Is: The Problem of Overlapping Hierarchies. Retrieved from [https://www.ideals.illinois.edu/handle/2142/9407], 2017/11/20.
- Seminara, S. (1998). *L’accadico di Emar*. Roma: Università degli Studi di Roma “La Sapienza”.
- Smith, E.J.M. (2007). Using LPath Queries to Annotate Corpora: A Case Study of Elamite and Sumerian. In P. Zemánek, J. Gippert, H.-C. Luschützky, & P. Vavroušek (Eds.), *Chatreššar 2007. Electronic Corpora of Ancient Languages. Proceedings of the International Conference Prague*,

*November 16-17, 2007* (pp. 121–134). Retrieved from [<http://usj.ff.cuni.cz/system/files/Smith-Ch-2007.pdf>], 2017/11/20.

Witt, A. & Metzger, D. (2010). *Linguistic Modeling of Information and Markup Languages: Contributions to Language Technology*. New York: Springer.

Yamada, M. (2013). The Chronology of the Emar Texts Reassessed. *Oriental*, 48, 125–156.

Christian Prager, Nikolai Grube, Maximilian Brodhun, Katja Diederichs, Franziska Diehr, Sven Gronemeyer and Elisabeth Wagner

## 5 The Digital Exploration of Maya Hieroglyphic Writing and Language

**Abstract:** The Maya hieroglyphic script (300 BCE–1500 CE), which has only been partially deciphered, is one of the most significant writing traditions of the ancient world. In 2014, the project *Text Database and Dictionary of Classic Mayan*<sup>1</sup> was established at the University of Bonn by the North Rhine-Westphalian Academy of Sciences, Humanities and Arts and the Union of the German Academies of Sciences and Humanities, to research the written language of the pre-Columbian Maya. The project aims to use digital methods and technologies to compile the epigraphic contents and object histories of all known hieroglyphic texts. Based on these data, a dictionary of the Classic Mayan language will be compiled and published near the end of the project's runtime in 2028. The project is methodologically situated in the digital humanities and conducted in cooperation with the Göttingen State and University Library (Grube & Prager, 2016).

**Keywords:** Maya hieroglyphic writing, digital epigraphy, virtual research environment, lexicography, XML/TEI

### 5.1 Introduction

The subject of our research project is the written language of the Classic Maya, whose cultural area extended over the territory of the present-day nation states of Mexico, Guatemala, Belize and Honduras (Figure 5.1). Maya writing was used for more than 1,500 years and can be found, for example, on free-standing monuments (stelae, altars), architectural elements (lintels, columns, door jambs), portable objects and in the natural environment, such as in caves or on rock faces (Grube, 2001). It has

---

1 Textdatenbank und Wörterbuch des Klassischen Maya [<http://mayawoerterbuch.de/>].

---

**Christian Prager, Nikolai Grube, Katja Diederichs, Sven Gronemeyer, Elisabeth Wagner**, Rheinische Friedrich-Wilhelms-Universität, Bonn

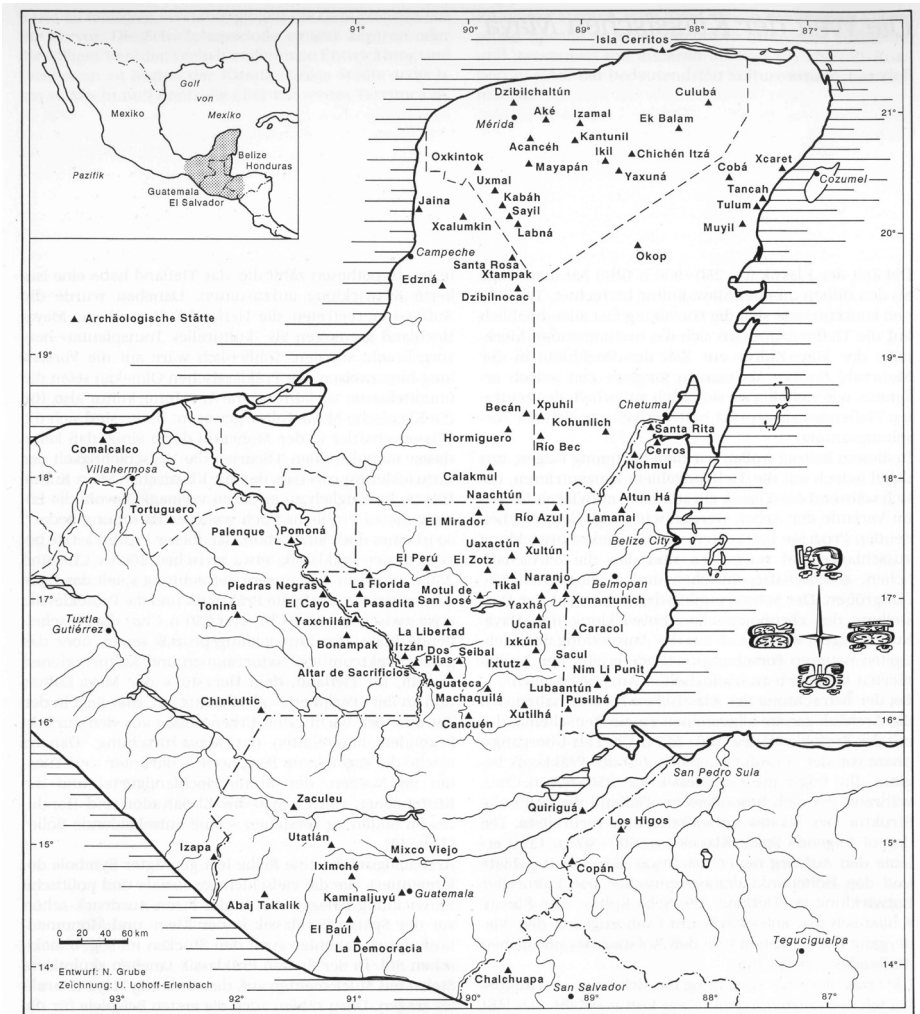
**Maximilian Brodhun, Franziska Diehr**, Niedersächsische Staats- und Universitätsbibliothek, Göttingen  
**Sven Gronemeyer**, La Trobe University, Melbourne



© 2018 Christian Prager *et al.*

This work is licensed under the Creative Commons Attribution 4.0 International license (CC BY 4.0)

survived on more than 10,000 text-bearing objects, dating between 300 BCE and 1500 CE and originating from more than 500 archaeological sites. The glottographic writing system comprises about 1,000 figurative graphs, most of which are signs for words or syllables. They represent figurative and abstract objects from the natural environment and material culture, human and animal body parts, heads of humans and animals or portraits of supernaturals, among other forms.



**Figure 5.1:** Map of the Yucatan peninsula with major archaeological sites (drawing by N. Grube and U. Lohoff-Erlenbach)

The language of the hieroglyphs, now called Classic Mayan, has been preserved in large part in colonial and modern Ch'olan and Yukatekan languages (Wichmann, 2006, p. 201). Many texts display calendar dates that record the exact sequence of events, providing unique data on the history of Maya writing and language. Classic Mayan can thus be reconstructed with chronological precision, and the results can be compared with findings from historical linguistics.

Many inscriptions originate from in or around the palaces of divine kings who ruled over independent city-states. The inscriptions often contain biographical information on political elites and provide written evidence for inter- and intra-dynastic connections between the ruling families.

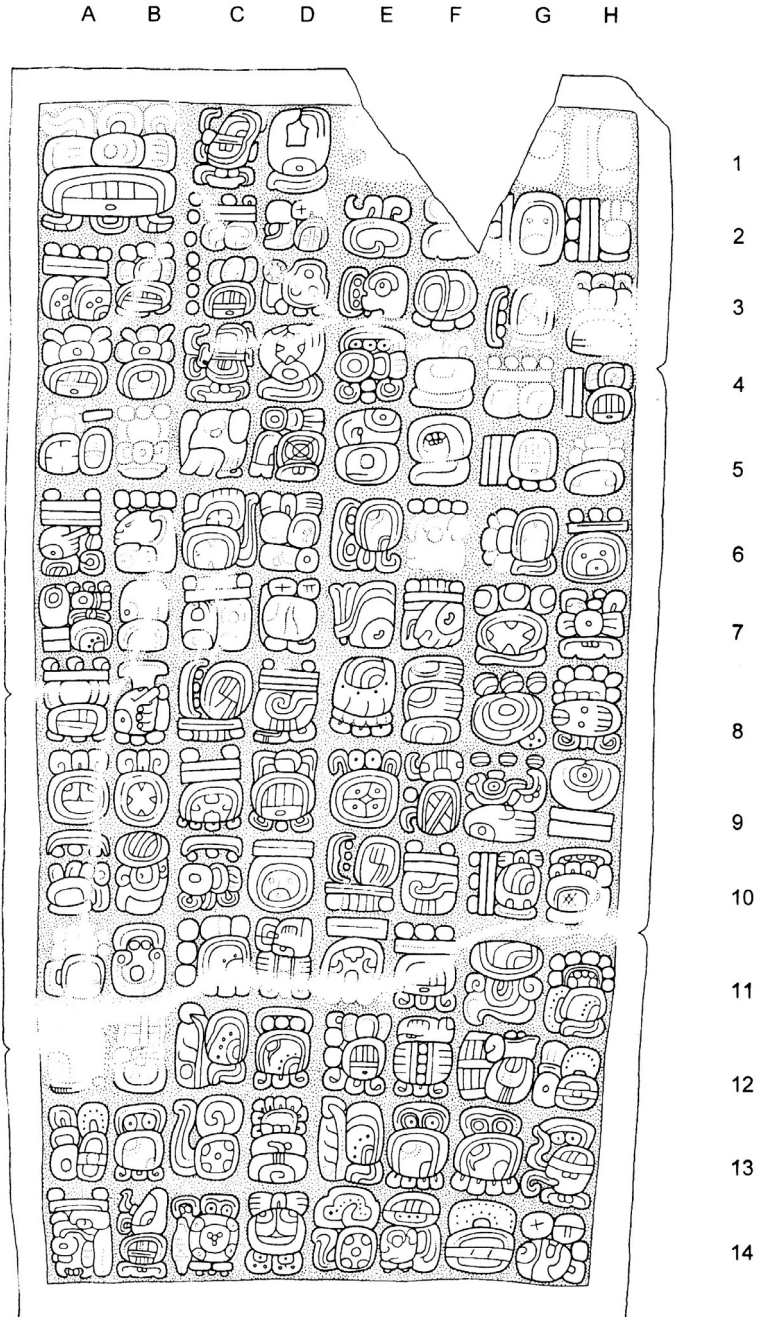
Some public monuments, like Stela D from Pusilha (Figure 5.2), describe actions such as war or royal visits. Others attest to ceremonies and religious rituals carried out in the context of accessions to the throne, ancestor worship, calendrical anniversaries, inaugurations, processions and other occasions that marked royal daily life (Martin & Grube, 2008).

## 5.2 Maya Hieroglyphic Writing

The Maya writing system is considered a hieroglyphic script because of the iconic character of its approximately constituent 1,000 graphemes. Typologically, it is a logosyllabic, or rather, a morphographic writing system with two basic, functional sign types: syllabic signs and morphographs. The latter denote concrete words and bound morphemes, whereas the former represent vowels and open syllables and thus permit the syllabic spelling of lexical and grammatical morphemes. In addition, syllabic signs were used as pre- or post-fixed phonetic complements for morphographs. Thus, it was possible to write words entirely with syllabic signs or by simply using morphographs.

Usually, however, morphographs and syllabographs were combined to form morpho-syllabic spellings of words (Figure 5.3). A high level of calligraphic complexity was further achieved through allographic notation and modification of graph shapes. More common syllables could be written with at least two or more graphemes, which explains the extremely high number of syllabic signs (about 300) relative to the total inventory of more or less 1,000 graphemes in the Maya script (Grube, 1994). This phenomenon allowed scribes to compose aesthetically ambitious texts that minimized sign repetition.

The signs were combined into roughly quadratic blocks (Figure 5.4), not unlike Korean Hangul. A single hieroglyphic block usually corresponds to the emic concept of a Classic Mayan word. In most texts, these blocks were arranged in double columns that were read from left to right and from top to bottom. Sentences were formed by sequencing hieroglyphic blocks to reflect various syntactic features, such as possession. Multiple sentences were joined to produce complex texts, whose syntax and discourse structure are comparable to those found in modern Mayan languages.



**Figure 5.2:** Stela D from the Maya site of Pusilha, Belize, with references to local dynastic and political history (drawing by C. Prager)



**Morphographs****Syllabographs***bahlam* „jaguar“*morphemic*

BALAM

*syllabic*

ba-la-ma

*morpho-syllabic*

BALAM-ma

**Figure 5.3:** Examples of basic sign functions in Maya writing (concept by C. Prager)

The individual elements within each hieroglyphic block are traditionally subdivided into main and small graphs; main graphs are spatially larger and approximately square in shape, whereas small graphs are attached to the periphery of the main characters and oriented along their vertical or horizontal axis. Within a block, individual graphs could be arranged side-by-side or on top of each other (affixation). They could also merge into a single graph (conflation). In addition, two or more graphs could partially or completely overlap (ligature), or one could be inserted into the other (infixation). Altering the shape of an individual graph or block had no influence on its pronunciation or meaning.

Graph morphology and the arrangement of glyphs into blocks are particularly challenging for epigraphers to interpret in those cases in which either all, or some, of the signs have not yet been deciphered, or have only been hypothetically deciphered and thus elude linguistic verification. Documenting the original spelling or graph arrangement using XML/TEI is therefore essential to epigraphic work with syllabic and morpho-syllabic hieroglyphic writing systems, since a simple, linear transcription of a text does not show original spellings or placement of the glyphs within the block (Prager & Gronemeyer, 2016).

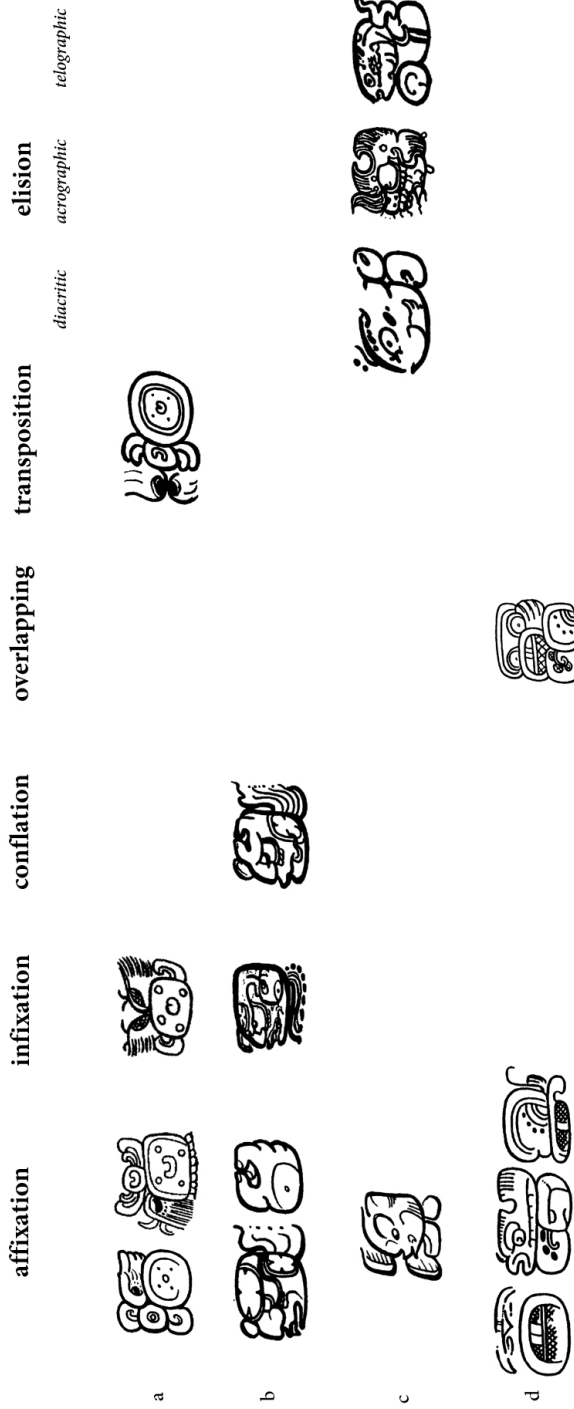


Figure 5.4: Graphotactic patterns in Maya writing. a) Syllabic spellings of *y=uk'ib* “his drinking vessel”. b) Different spellings of *K'inich*, the proper name of the Classic Maya sun god. c) Scribal plays for the word *kakaw* “cacao”. d) Two spellings of the word *u=pakbu tuunil* “his stone-lintel”

The aforementioned graphemic and graphotactic strategies affected only the graphic realization of words in the Maya script. The principle of underrepresenting specific word-endings, in contrast, impacted both visual form and pronunciation of the hieroglyphs. Omission through underrepresentation enabled scribes to graphically vary individual words and texts. This scribal practice also had an impact on the Classic Mayan lexicon, because underrepresenting phonemes in writing elicits different pronunciations of a given word, which have to be considered and examined in the context of this dictionary project. Using this wide range of graphemic and graphotactic strategies, Maya scribes were able to create a wide variety of texts that avoided repeating the same graphs or spellings. This technique suggests that these artists sought to maximize visual splendor and designed texts and pictorial works as individual pieces, even though their contents are often rather formulaic and stereotypical (Zender, 1999).

### 5.2.1 Decipherment

Considerable breakthroughs have already been achieved in the decipherment of the Classic Mayan written language (Houston & Martin, 2016). However, despite the great progress made in recent decades, some 30% of the script's 1,000 signs remain unreadable, even today. One reason is their lack of systematic attestation. Even in cases in which individual signs are legible, texts may still elude understanding because the Classic Mayan language itself has not survived; instead, it can only be reconstructed through historical linguistic comparison of the 30-odd Mayan languages that have been documented since European conquest, most of which are still spoken today (Wichmann, 2006). However, much pre-Hispanic Mayan vocabulary has been lost in the aftermath of European colonization. Consequently, comprehensive documentation and decipherment of the approximately 10,000 extant hieroglyphic texts, reconstruction of the language that they record, and documentation of that language in a dictionary, are necessary to acquire a deeper understanding of Classic Maya culture, history, religion and society.

Recent research on Maya writing and language has addressed material form and alternative reading hypotheses for graphemes with varying degrees of plausibility, as well as vague semantic interpretations of hieroglyphs and text passages. For many graphemes, multiple readings have been proposed whose plausibility we evaluate by means of propositional logic based on individual arguments for decipherment that have been published in the literature. In this manner, we can qualify proposed readings and decipherments, enabling us to distinguish in the text database and dictionary between hypothetical readings and secure decipherments. By using propositional logic, for example, our evaluation of the more than six readings that have been proposed for the “star war” hieroglyph, which expresses war against a given site (see Stuart, 1995; Martin, 1996; Aldana, 2005; Chinchilla Mazariegos,

2006; Voit, 2013; also Macri &Looper, 2003; Macri & Vail, 2009), reveals that David Stuart's (1995, p. 313) proposal, according to which the glyph represents a logographic substitution JUB "fall" for the attested syllabic spelling **ju-bu**, seems to be the most promising candidate. A detailed discussion of our propositional logic can be found in section 3.2.3.

Modelled in this way, our text database and dictionary of Classic Mayan will not only represent the results of our research; in addition, both components will serve as tools for further studying Classic Maya writing and language. On the one hand, our database architecture is designed to reflect the dynamics and processuality of Maya hieroglyphic research. On the other, ongoing integration of new research results will permit us to continuously improve the quality of our data. It is also important for our database work to consider the relation between text and context: hieroglyphic inscriptions very often refer to the text-bearing object itself, including its spatial, temporal and social context. When compiling the dictionary or analysing the meaning of words, Classic Mayan texts should not be considered independently of the object on which they are recorded, nor of their temporal or spatial context. The object and its context provide non-textual information, or metadata, about the text-bearing object itself, its location, neighboring texts and associated finds, its commissioner, and its historical context as a whole. These data are highly significant for deciphering and interpreting the inscriptions, and carefully documenting them in the database is a prerequisite for successful decipherment and text interpretation (Prager, 2015).

### 5.2.2 Sign Lists and Classification

Since roughly one-third of the script's signs still cannot be read, decipherment of Classic Mayan remains a frequently discussed research topic that has generated a variety of hypotheses about possible sign readings. To address the challenge of discussing signs with no known reading, epigraphers have established different inventories that assign each graph an (alpha)numeric value (Zimmermann, 1956; Thompson, 1962; Grube, 1990). Thus, a descriptive transliteration can be given independently of the signs' (different possible) phonemic values. We also aim to employ a numeric transliteration as a basis for the digital mark-up of Maya writing.

The first step is to develop a digital inventory of Maya signs and graphs. As noted in the discussion of Classic Maya graphemics above, many graphs in the script can have several variants. Yet, to this day there is no complete inventory or classification of all graphs. Identifying and cataloguing individual graphs and allographs thus represent central challenges for our project. For this reason, we first systematically reviewed existing catalogues and, eventually, opted for a modelling approach in which each graph is recorded separately from the corresponding sign's phonemic representation. Thus, we can exactly document individual graph variants, but also establish relations to other graphs to point out common diagnostic features. Uniquely,

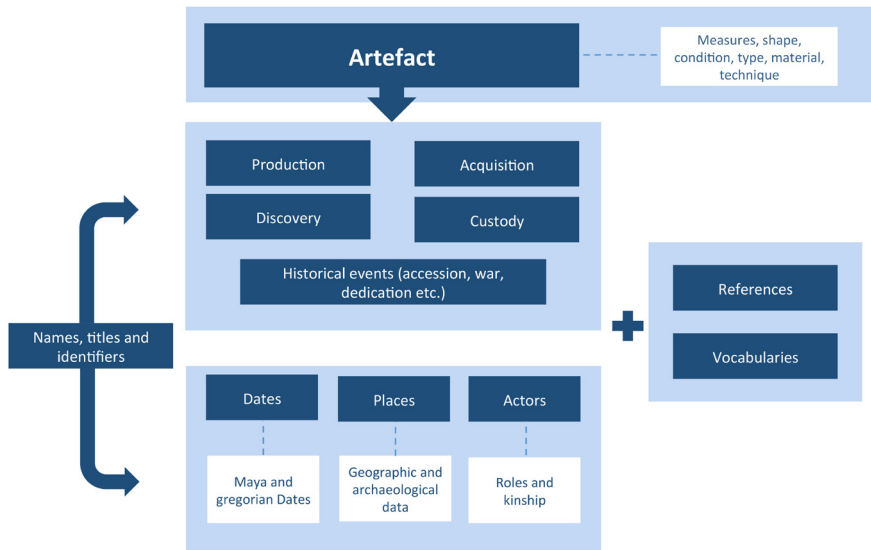
our sign catalogue records signs and their graph realizations as separate entities, with each receiving catalogue numbers and URIs. As such, we can be flexible in classifying the sign-grapheme relation. Each sign can be assigned multiple functions and thus also multiple transliteration values, and each can be related to 0 - n graphs (Diehr et al., 2017).

### 5.3 Digital Epigraphy of Classic Mayan

For digital documentation and epigraphic analysis of the text-bearing objects, we distinguish between object documentation in RDF, representation in the sign and graph database, and linguistic analysis. This work will make use of the Java-based multi-level annotation tool ALMAH (Annotator for the Linguistic Analysis of Maya Hieroglyphs), which is currently being developed in cooperation with Cristina Vertan of the University of Hamburg. In this section, we discuss these areas of work and our central concerns when creating the virtual research environment for digitally investigating Maya writing.

#### 5.3.1 Documentation of Object Information

Documenting and recording text-bearing objects is the foundation of the project's textual analysis. Consequently, the project began by designing and subsequently constructing its information technology infrastructure. As noted previously, the material form in which Classic Mayan texts are recorded, and their temporal or spatial context, provide metadata about the text-bearing object and its sociocultural context. Because these data are critical to deciphering and interpreting the inscriptions, future decipherment and text interpretation depends on carefully documenting them in the database. In addition to linking with the text database, the object database can also establish and query relationships between multiple texts and text-bearing objects. This is made possible by its ontology-based modelling and implementation in a RDF-data model. Figure 5.5 gives an overview of the data structure and shows how relations can be established between artifacts, events, dates, places, appellations, references and vocabularies. Furthermore, the database also connects to the literature database compiled in Zotero. With this feature, every unit of information recorded about the text-bearing object (such as date, events, persons, its measurements, artefact type, shape, condition, find-spot, archaeological context, etc.) can be referenced with a bibliographic citation. In this way, the user can acquire an overview of who has studied or published about a monument or has discussed a text passage (Diederichs et al., 2016).



**Figure 5.5:** Overview of the ontology-based metadata schema for describing artefacts and their contexts

### 5.3.1.1 Controlled Vocabularies

The project has developed a total of 10 multilingual thesauri. In choosing appropriate entries, we prioritized normed data, such as the Getty Research Institute’s Art & Architecture Thesaurus (AAT),<sup>2</sup> and the Getty Thesaurus of Geographic Names (TGN).<sup>3</sup> Since the Mesoamerican, and particularly the Maya cultural spheres, are still underrepresented in the Getty thesauri, we checked a significant number of terms that had been previously employed in the literature for plausibility, comparability, and utility by for instance, consulting specific encyclopediae (e.g., Loten & Pendergast, 1984; Gendrop, 1997; Witschey, 2016), monument corpora (e.g., Graham, 1975; Jones & Satterthwaite, 1982), and archaeological reports (e.g., Culbert, 1993).

The resultant collection of terms was ordered according to terminological principles and modelled in the SKOS (Simple Knowledge Organization System) format in order that they could be represented in machine-readable format and

<sup>2</sup> Getty Research Institute’s Art & Architecture Thesaurus (AAT) [<http://www.getty.edu/research/tools/vocabularies/aat/index.html>].

<sup>3</sup> Getty Thesaurus of Geographic Names (TGN) [<http://www.getty.edu/research/tools/vocabularies/tgn/index.html>].

integrated into the metadata schema.<sup>4</sup> The terms could thus also be simultaneously mapped onto normed data from the Getty Thesaurus, allowing the reused terms to be referenced. A special feature of SKOS is that vocabularies are represented in a concept-based manner, i.e., there are concepts with several labels, whereby one preferred label exists for each concept, to which any number of other labels can be added as alternative or hidden labels. Over the course of more than 150 years of Maya research, numerous alternative denominations for objects, persons or place names have become established in the literature, which we will document and further differentiate into preferred and alternative labels. This feature, which is useful for clarifying nomenclature and providing insight into the history of the field, also supports our work by identifying alternative or obsolete terms in the literature (Grube et al., 2016).

Developing the controlled vocabularies is highly beneficial not only to the project's own work, but also to the discipline more broadly. Until now, a multitude of terms, vocabularies, and descriptive schemas has existed in Maya epigraphy, resulting in a wide range of differentially documented text-bearing objects. At times, records exhibit relatively little agreement in application of existing terminology and are often incomplete, erroneous, imprecise, or dramatically simplified. In developing these vocabularies, the project is making a significant contribution to terminological standardization in Maya epigraphy, because we reuse terms that are already established in other scientific fields, but clearly define them for the first time and situate them in a terminological relationship to one another.

### 5.3.1.2 Technical Infrastructure

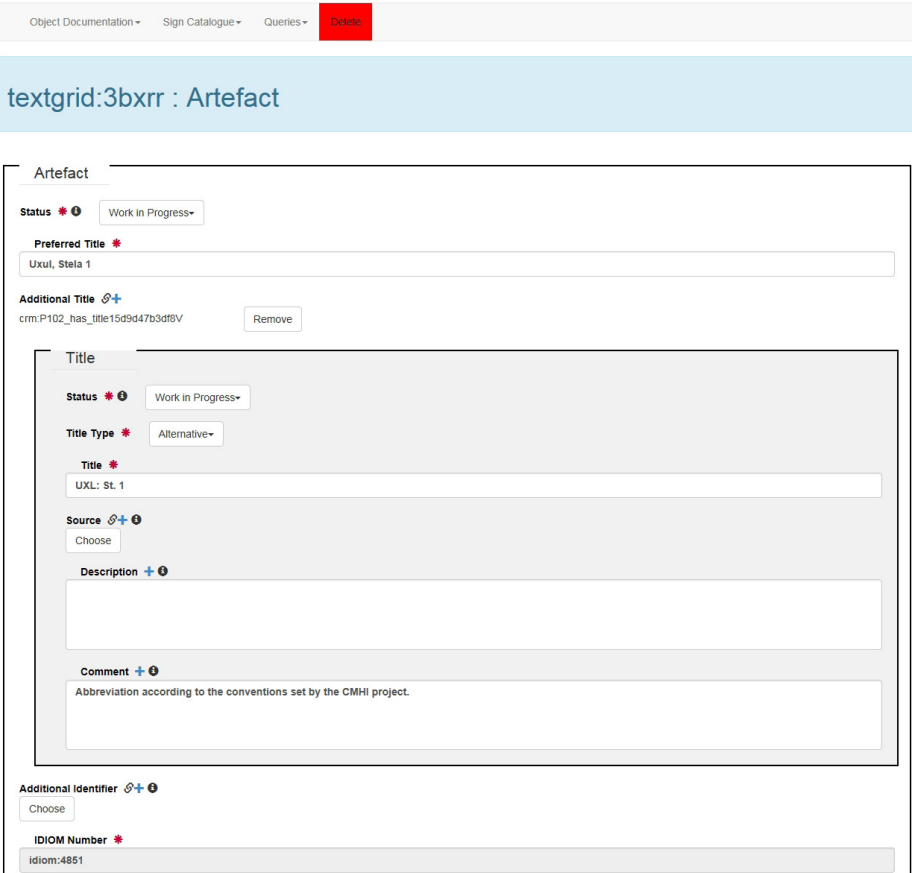
Data of various types are being created and stored as part of the project's workflow. Image data, metadata, and text analysis files must be managed in relation to one another within a single infrastructure for creation, storage, processing, and access regulation. We are using the virtual research environment TextGrid for these tasks. The front-end TextGrid Laboratory (TG Lab) allows files to be created and processed, in addition to facilitating fine-grained management of the rights thereto. The back-end provides access to the repository (TG Rep), which stores the data in a secure environment.

The densely networked structure of the files with metadata for text-bearing objects requires appropriate storage, which we achieve by using the format RDF (Resource Description Framework) and by storing them in a graph database in the form of a triple store. An entry mask is used to record the metadata in a user-friendly manner (Figure 5.6). This HTML- and JavaScript-based tool provides the user with multiple entry aids. When utilized as a plug-in, the entry mask can be installed and directly

---

<sup>4</sup> The metadata schema can be retrieved from [idiom-projekt.de/idiommask/schema.html].

used from the TG Lab. Examples of its supporting functions include searches in internal and external databases to establish object relations, validating entry fields, and automatically converting data formats (Neuroth, Rapp, & Söring, 2015).



Object Documentation - Sign Catalogue - Queries - **Delete**

textgrid:3bxrr : Artefact

Artefact

Status \*

Preferred Title \*

Additional Title

Title

Status \*

Title Type \*

Title \*

Source

Description

Comment

Additional Identifier

IDIOM Number \*

Figure 5.6: HTML and JavaScript-based entry mask to record the metadata in TextGrid

### 5.3.2 Documentation of Signs and Graphs

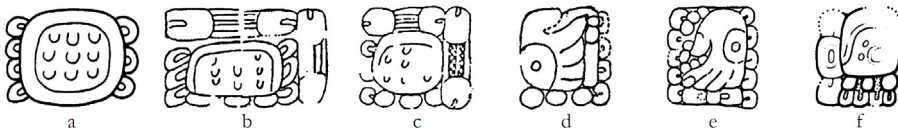
Documenting the original presentation of each hieroglyphic text is essential to our epigraphic work, because a linear transliteration and transcription does not display the original spellings or arrangement of the hieroglyphic inscription. Thus, texts are annotated at the level of the graph, using numeric values to represent each graph. Linguistically speaking, the graph, as the smallest graphic unit of a writing system, has not yet been assigned to a grapheme, a sign representing a phonemic value. In



our virtual research environment, each graph is linked to the digital sign catalogue, which is a prerequisite and starting point for epigraphic text analysis (Diehr et al., 2017).

### 5.3.2.1 Modelling Graph Variants

The concept for the digital sign catalogue requires a data structure that enables formation of semantic relations between clearly referenced entities. As with the metadata organization for documenting object information, a data model implemented in RDF represents the optimal form of knowledge representation, and we opted to use the CIDOC Conceptual Reference Model (CRM) as the basic ontology. CIDOC CRM contains many appropriate meta-concepts that are suitable for structuring our digital sign catalogue. According to this model, a graph can be related to the functional and phonemic level of representation (modelled as the class *Sign*). This linking is optional, so that we can also register graphs that cannot yet be assigned to any sign.



**Figure 5.7:** Allographic spellings of the sign 595 for the syllable *no*. a) Full form of 595, represented by the graph 595tv. b-c) Sign 595 in the spelling *ko-ko=no=ma* < *kok-n-om* “guardian” d) 595 used in the word *TZUTZ=no=ma* < *tzutz-n-om* “planter”, e) *CHOK=no=ma* < *chok-n-om* “scatterer”, f) *yu-ku=no=ma* < *yuk-n-om* “shaker” (drawings by Stephen Houston, Linda Schele)

Here, a catalogue number is assigned to each sign, based on Eric Thompson’s catalogue of Maya hieroglyphs (Thompson, 1962). For instance, sign number 595 represents the syllable **no** (Figure 5.7). When a graph has been identified as an allograph, it is assigned a *graphNumber* consisting of the catalogue number of the character and the abbreviation of the variation type (e.g., 595tl, where tl = tripartite left, meaning that the sign 595 is represented by the left-hand segment of a graph that can be cut in three vertical segments) (Diehr et al., 2017; Prager & Gronemeyer, 2016).

### 5.3.2.2 Modelling Multiple Sign Functions

A feature of Maya writing is that a sign can have several readings or sign functions. The sign denoted with 528, for example, can be read as the morphograph **TUN** “stone”, as the morphograph **CHAHUK** for the name of a Maya day, or as the syllable **ku** (Figure 5.8).



**Figure 5.8:** Graphs of sign 528 representing the syllable ku, the morphograph TUN “stone”, and the day sign CHAHUK, the name of 19th day in the Maya calendar (drawings by M. Zender)

The graphs themselves do not indicate which reading or sign function is intended. Therefore, in our metadata schema, we consider the following possible sign functions: numerals, diacritical signs, morphographs with identified linguistic reading, morphographs with unidentified reading (in which case a meaning is assigned to the sign) and syllabic signs with identified reading. To represent them in the schema, we have modelled the class *SignFunction*, with the aforementioned functions as subclasses. The reading, or rather the transliteration value, is recorded as the corresponding sign function (Diehr et al., 2017).

### 5.3.2.3 Evaluating Sign Readings

Undeciphered signs inspire lively discourse in Maya hieroglyphic research, from which new proposals for their linguistic decipherment are constantly emerging. Examples include discussions of recent reading proposals on David Stuart’s specialist blog, short articles journal *Mexicon*, or publications on our project’s website.<sup>5</sup> New reading hypotheses must therefore be integrated into the digital sign catalogue so that they can be analyzed in the corpus and evaluated for plausibility. In order to formally assess the quality of each linguistic decipherment, we have developed a set of criteria for sign function that are based on the linguistic context of use (e.g., part of speech, plausible text-picture reference, etc.) or lexical evidence from modern Mayan languages. The criteria for decipherment are related by means of propositional logic that produces a quality level for each reading proposal depending on its particular combination. Figure 5.9 shows the evaluation of sign 528 and its transliteration value as the morphograph TUN. To represent these evaluations in the Sign Catalogue, the class *ConfidenceLevel* was modelled, which is placed in relation to the class *SignFunction*. Therefore, a qualitative rating of the reading confidence can be obtained for each transliteration value recorded as the sign function. This rating is particularly relevant for determining the plausibility of a reading proposal within the text corpus. For example, linguistic decipherments with a particularly high level can be compared

<sup>5</sup> David Stuart’s Blog “Maya Decipherment”: [<https://decipherment.wordpress.com/>]; *Mexicon* - The Journal of Mesoamerican Studies: [[www.mexicon.de](http://www.mexicon.de)]; The project’s website: [[www.mayadictionary.de](http://www.mayadictionary.de)].

with those with a low level. For readings with a low confidence level, new criteria for their plausibility could also be found through later research with hieroglyphic sources and the entries can then be updated in the digital Sign Catalogue (Diehr et al., 2017).



### Sign No. 528 with transliteration value **TUN**

selected parameter	rules for reasoning
✓ <b>k</b> complete phonetic substitution	1 $d \vee h \vee (k \wedge p)$
✓ <b>c</b> postponed complementation [-suffix(es)]	2 $((o \vee y) \wedge c) \vee t$
✓ <b>u</b> preconsonantal ergative pronoun	3 $U \wedge c \wedge (l \vee m \vee n) \wedge s \wedge i$
✓ <b>p</b> expected part of speech	4 $(o \vee c) \wedge (l \vee m \vee n) \wedge s$
✓ <b>l</b> attested in GL languages	5 $g \wedge (l \vee m \vee n) \wedge p \wedge s$
✓ <b>m</b> attested in other Mayan languages	6 $g \wedge s \wedge p$
✓ <b>i</b> semantic correspondence with graph icon	7 $i \wedge s \wedge (u \vee y)$
✓ <b>s</b> semantic correspondence with context	8 $i \wedge s$
<b>k + c + u + p + l + m + i + s</b>	

$$d \vee h \vee (k \wedge p) = \text{level 1}$$

**Figure 5.9:** Example evaluation of the transliteration value “TUN” for Sign No. 528

#### 5.3.2.4 Components for Generating a Digital Corpus

To create a machine-readable text corpus, there must be a text that can be encoded. For Maya hieroglyphic writing, we are confronted with the problem that, due to the complexities of calligraphy and text arrangement, we cannot use a standardized font such as Unicode. Secondly, signs may fulfil several functions with various proposed readings, as explained above. Therefore, we cannot encode phonemic-transliterated values, as this would preclude tagging graph variants. However, as we intend to study these variants and their usage, marking-up graphs and allographs is necessary to generate the digital corpus of Maya hieroglyphic inscriptions. Therefore, the only possibility for creating a machine-readable text is to refer to the graphic representation itself, which is where the digital Sign Catalogue comes into play: in the TEI/XML encoding process, each glyph is recorded by referring to the URI of the graph recorded in the digital Sign Catalogue. A subsequent processing step produces a human-readable text that is enriched with the transliteration values stored in the digital Sign Catalogue. On this basis, linguistic analyses can be conducted that account for the multiple functions and the various reading suggested for each sign (Diehr et al., 2017).

### 5.3.2.5 A TEI Schema for Digitally Documenting Maya Inscriptions

One of the project's central tasks is to develop a metadata schema for documenting Maya hieroglyphic inscriptions. As we have outlined above, documenting the original spelling is a fundamental aspect of epigraphic work with syllabic and morpho-syllabic hieroglyphic writing systems, since mere transliteration and transcription do not represent the original spelling. To digitally document Maya texts, we use XML/TEI. In XML/TEI, texts are annotated at graph level using numeric values to represent the graphs. Within the XML/TEI document, each graph is linked to the digital sign catalogue with URIs. To encode the graphotactics of Maya hieroglyphs and represent the original spelling and graphic arrangement of each glyph block, we specified the attribute values @rend in TEI element <g> to indicate the spatial relation among single graphs. Furthermore, using TEI allows us to record information concerning text structure or conservation status, graph colour, shape and location of the text field, and so on.

Our aim is to digitally represent all features of text arrangement. Thus, an important requirement for the TEI schema is that it digitally represents the semantic and topographic structure of a hieroglyphic text and its associated iconography. The schema must display the logical reading order, as well as the actual text arrangement and graph order. Topographically, this means that the schema must indicate the text's location and the position of each graph in relation to its neighbours. The semantic text structure should show how a hieroglyphic text is read and of which logical sequence it consists.

In another step, we also addressed the question of how to deal with unreadable, vague or reconstructed text passages and how to model them in the TEI schema. Our approach is to develop a text-critical analysis that can be taken into account in epigraphic and linguistic analysis of hieroglyphic inscriptions. Thus, our TEI schema also accommodates unclear or restored text passages that have been damaged or destroyed by physical, chemical, or biological influences. We also attend to text design, i.e., we characterize the inscription's design and typography, as well as the relationship between text and image: what criteria for text design may be relevant to our research questions, and what is the relationship between design and semantics? Our TEI schema therefore records characteristics such as the form of a text field, relief, framing, coloration, or font size, as well as individual hands of scribes or workshops (Maier, 2015; Diederichs et al., 2016; Diehr et al., 2017).

### 5.3.2.6 Multi-Level, Semi-Automatic Annotation of Classic Mayan

The TEI markup still lacks any linguistic annotation and analysis. For this, we are cooperating with Cristina Vertan to use an XML-based tool, ALMAH (Annotator for the Linguistic Analysis of Maya Hieroglyphs), originally developed for semi-

automatic annotation of *fidal*, the Old Ethiopic script.<sup>6</sup> This analysis tool will be adapted for epigraphic and linguistic analysis of Maya hieroglyphic texts to allow semi-automatic, multi-level annotation. ALMAH will allow us to create dictionary entries from analyses of hieroglyphic inscriptions. Just like the digital sign catalogue, this tool will be adaptable so that it can continually incorporate new research findings about Classic Mayan grammar and morpho-syntax. In contrast to traditional epigraphic analysis that focuses on transliterating and transcribing, this approach will include steps that not only reflect the need to achieve machine-readability in a granular and transparent way, but also increase comprehensibility of analysis in general. Transparency in analysis, accommodation of incomplete decipherments, integration of reading hypotheses and connections to the object data schema and the data contained therein: these features constitute our digital approach to not only compiling a semi-deciphered writing system and language in a dictionary, but also to deciphering them in the near future.

## 5.4 Summary and Conclusion

The subject of the project *Text Database and Dictionary of Classic Mayan* is an incompletely deciphered, complex writing system. The project aims to decipher it using digital tools and will describe its underlying language in a dictionary. To these ends, Maya hieroglyphic texts are being made machine-readable using XML/TEI and saved in a text database with analysis and commentary. In addition, the Classic Mayan language is represented in its original orthography in a web-based dictionary, which will allow users to compare the content with its analysis. This is a desideratum that we can also identify in the study of other ancient writing systems. The documentation of original spellings and references to the entire text has often been lacking in Egyptology, for example, where standardized representation of hieroglyphs is the norm. The digital age can easily remedy this shortcoming.

Even a glance at the epigraphic projects united in this volume indicates that Maya epigraphy is not alone in confronting the challenges presented by complex, hieroglyphic and morpho-syllabic writing systems, as exemplified by the Sinleqiunnini, OIMEA or HPM projects. However, when developing databases, most research projects in digital epigraphy do not usually face the additional difficulty of their respective writing systems and corresponding languages being only partially, or not at all, deciphered. Our goal is to use digital tools to compile and register newly classified signs in sign lists, make the texts machine-readable, discern readings, and document the Classic Mayan vocabulary in its original representation.

---

<sup>6</sup> Vertan, Cristina. GeTa, a multi-level semi-automatic annotation tool for Classical Ethiopic. DOI: <https://doi.org/10.5281/zenodo.160366>

The project's outcomes will ultimately include developing tools, methods and standards for digital research on ancient writing systems and for the digital humanities as a whole, in addition to producing content about the Maya script. The project's emphases on digital epigraphy, knowledge representation, database development and long-term and interoperable storage of research data, in particular, underscore the great significance of the digital humanities for such an innovative undertaking. Yet, we are also contributing to computer-based research on writing systems and developing methods and standards that will benefit other areas of research.

## Bibliography

- Aldana, G. (2005). Agency and the "Star War" Glyph: A Historical Reassessment of Classic Maya Astrology and Warfare. *Ancient Mesoamerica*, 16(2), 305–320.
- Chinchilla Mazariégoz, O. (2006). *A Reading for the "Earth-Star" Verb in Ancient Maya Writing* (Research Reports on Ancient Maya Writing 56). Barnardsville, NC: Center for Maya Research.
- Culbert, T.P. (1993). *The Ceramics of Tikal—Vessels from the Burials, Caches and Problematical Deposits* (Tikal Report 25A. University Museum Monograph 81). Philadelphia, PA: University of Pennsylvania Museum of Archaeology and Anthropology.
- Diederichs, K., Gronemeyer, S., Prager, C., Wagner, E., Diehr, F., Brodhun, M., & Grube, N. (2016). A Virtual Research Environment to Document and Analyze Non-alphabetic Writing Systems: A Case Study for Maya Writing. In S. Orlandi, R. Santucci, F. Mambrini, & P.M. Liuzzo, (Eds.), *Digital and Traditional Epigraphy in Context. Proceedings of the EAGLE 2016 International Conference* (pp. 233–246). Roma: Sapienza Università Editrice. doi: 10.13133/978-88-9377-021-7
- Diehr, F., Brodhun, M., Gronemeyer, S., Diederichs, K., Prager, C., Wagner, E., & Grube, N. (2017). Modellierung eines digitalen Zeichenkatalogs für die Hieroglyphen des Klassischen Maya. In M. Eibl & M. Gaedke (Eds.), *Informatik 2017* (pp. 1185–1196). Bonn: Gesellschaft für Informatik. Retrieved from [<https://dl.gi.de/handle/20.500.12116/3882>], 2017/12/08.
- Gendrop, P. (1997). *Diccionario de Arquitectura Mesoamericana*. Mexico-City: Trillas.
- Graham, I. (1975). *Corpus of Maya Hieroglyphic Inscriptions. Vol. 1.1: Introduction*. Cambridge, MA: Peabody Museum Press.
- Grube, N. (1990). *Die Entwicklung der Mayaschrift: Grundlagen zur Erforschung des Wandels der Mayaschrift von der Protoklassik bis zur spanischen Eroberung* (Acta Mesoamericana). Berlin: Von Flemming.
- Grube, N. (1994). Observations on the History of Maya Hieroglyphic Writing. In V.M. Fields (Ed.), *Seventh Palenque Round Table, 1989* (The Palenque Round Table Series) (pp. 177–186). San Francisco, CA: Pre-Columbian Art Research Institute.
- Grube, N. (2001). Hieroglyphs, the Gateway to History. In N. Grube, E. Eggebrecht, & M. Seidel (Eds.), *Maya: Divine Kings of the Rain Forest* (pp. 115–127). Köln: Könemann.
- Grube, N. & Prager, C.M. (2016). Vom Regenwald ins World Wide Web. In Union der deutschen Akademien der Wissenschaften (Ed.), *Die Wissenschaftsakademien - Wissensspeicher für die Zukunft: Forschungsprojekte im Akademienprogramm* (pp. 16–17). Berlin: Union der deutschen Akademien der Wissenschaften.
- Grube, N., Prager, C., Diederichs, K., Gronemeyer, S., Wagner, E., Brodhun, M., & Diehr, F. (2016). Annual Report for 2015. *Textdatenbank und Wörterbuch des Klassischen Maya*. doi: 10.20376/IDIOM-23665556.16.pr003.en

- Houston, S.D. & Martin, S. (2016). Through Seeing Stones: Maya Epigraphy as a Mature Discipline. *Antiquity*, 90(350), 443–455.
- Jones, C. & Satterthwaite, L. (1982). The Monuments and Inscriptions of Tikal—The Carved *Monuments*. (Tikal Report 33A. University Museum Monograph 44). Philadelphia, PA: University of Pennsylvania Museum of Archaeology and Anthropology.
- Loten, H.S. & Pendergast, D.M. (1984). *A Lexicon for Maya Architecture*. Toronto: Royal Ontario Museum. Retrieved from [<https://archive.org/details/lexiconformayaar00lote>], 2018/02/15.
- Macri, M.J. & Looper, M. (2003). *The New Catalog of Maya Hieroglyphs: The Classic Period Inscriptions* (Civilization of the American Indian Series 247). Norman, OK: University of Oklahoma Press.
- Macri, M.J. & Vail, G. (2009). *The New Catalogue of Maya Hieroglyphs: The Codical Texts* (Civilization of the American Indian Series 264). Norman, OK: University of Oklahoma Press.
- Maier, P. (2015). Ein TEI-Metadatenchema für die Auszeichnung des Klassischen Maya. *Textdatenbank und Wörterbuch des Klassischen Maya*. doi: 10.20376/IDIOM-23665556.15.wp003.de
- Martin, S. (1996). Tikal's "Star War" Against Naranjo. In M.J. Macri & J. McHargue (Eds.), *Eighth Palenque Round Table, 1993* (Palenque Round Table Series 10) (pp. 223–236). San Francisco, CA: Pre-Columbian Art Research Institute.
- Martin, S. & Grube, N. (2008). *Chronicle of the Maya Kings and Queens: Deciphering the Dynasties of the Ancient Maya* (2<sup>nd</sup> ed.). London: Thames & Hudson.
- Neuroth, H., Rapp, A., & Söring, S. (Eds.). (2015). *TextGrid: Von der Community - für die Community: eine virtuelle Forschungsumgebung für die Geisteswissenschaften*. Glückstadt: Hülsbusch.
- Prager, C.M. (2015). Das Textdatenbank- und Wörterbuchprojekt des Klassischen Maya: Möglichkeiten und Herausforderungen digitaler Epigraphik. In H. Neuroth, A. Rapp, & S. Söring (Eds.), *TextGrid: Von der Community - für die Community: Eine Virtuelle Forschungsumgebung für die Geisteswissenschaften* (pp. 105–124). Glückstadt: Werner Hülsbusch.
- Prager, C.M. & Gronemeyer, S. (2016, in press). *Neue Ergebnisse in der Erforschung der Graphemik und Graphetik des Klassischen Maya*. Retrieved from [[https://www.academia.edu/33672448/Neue\\_Ergebnisse\\_in\\_der\\_Erforschung\\_der\\_Graphemik\\_und\\_Graphetik\\_des\\_Klassischen\\_Maya](https://www.academia.edu/33672448/Neue_Ergebnisse_in_der_Erforschung_der_Graphemik_und_Graphetik_des_Klassischen_Maya)], 2017/12/08.
- Stuart, D. (1995). *A Study of Maya Inscriptions* (PhD Dissertation). Nashville, TN: Vanderbilt University.
- Thompson, J.E.S. (1962). *A Catalogue of Maya Hieroglyphs* (The Civilization of the American Indian Series). Norman, OK: University of Oklahoma Press.
- Voit, C.A. (2013). *The Venus "Shell-Over-Star" Hieroglyph And Maya Warfare: An Examination Of The Interpretation Of A Mayan Symbol* (MA Thesis). Detroit, MI: Wayne State University.
- Wichmann, S. (2006). Mayan Historical Linguistics and Epigraphy: A New Synthesis. *Annual Review of Anthropology*, 35, 279–294.
- Witschey, W.R.T. (Ed.). (2016). *Encyclopedia of the Ancient Maya*. Lanham, MD: Rowman & Littlefield.
- Zender, M. (1999). *Diacritical Marks and Underspelling in the Classic Maya Script: Implications for Decipherment* (MA Thesis). Calgary: Department of Archaeology, University of Calgary.
- Zimmermann, G. (1956). *Die Hieroglyphen der Maya-Handschriften* (Abhandlungen aus dem Gebiet der Auslandskunde. Reihe B, Völkerkunde, Kunstgeschichte und Sprachen 62). Hamburg: Cram, de Gruyter & Co.

Alessandro Bausi and Pietro M. Liuzzo

## 6 Inscriptions from Ethiopia. Encoding Inscriptions in Beta Maṣāḥəft

**Abstract:** This paper describes the available corpus of inscriptions from the Ethiopian and Eritrean regions giving an overview of this documentation. Some of the challenges involved with the inclusion of these documents in the Beta Maṣāḥəft project are presented: the connection to already digitally encoded texts, the encoding of the parallel *fidal* (i.e. Ethiopian script) and transcribed text, and the structuring of the data for the pseudo-trilingual inscription *RIÉ* nos 185 and 270 (that also has a second copy).

**Keywords:** Ethiopia, Eritrea, epigraphy, EpiDoc, multiple copies

### 6.1 Ethiopian and Eritrean Ancient Epigraphy

The Ethiopian and Eritrean region, despite the small numbers of inscriptions (amounting to some hundreds), offers examples for several case studies and a wide variety of languages and material epigraphic typologies. A rough estimation of the available inscriptions, arranged in chronological order and with obvious overlapping, including the few produced in Ethiopian languages or by Ethiopians in Antiquity, Late Antiquity and Middle Ages outside of Ethiopia (Yemen, Sudan, and Egypt), gives the rough figures detailed below.<sup>1</sup> Some of the entry numbers in the classical collection *Recueil des inscriptions de l'Éthiopie des périodes préaxoumite et axoumite* (*RIÉ*; Bernard, Drewes, & Schneider, 1991) are distinguished by an additional mark. In a few cases, one number refers to more inscriptions. This is typically the case of the two sets of the so-called “pseudo-trilingual” royal inscriptions, *RIÉ* nos 185 and 270, which correspond to two sets of three parallel texts each, respectively in Ethiopic in South Arabian script (*RIÉ* nos 185 I and 185bis I), in Ethiopic in Ethiopic non-vocalized script (*RIÉ* nos 185 II and 185bis II), and in Greek language and script (*RIÉ* nos 270 and 270bis). All in all, the two entries *RIÉ* nos 185 and 270 include six distinct inscriptions. Moreover, after the publication of *RIÉ* in 1991, additional inscriptions were discovered

---

<sup>1</sup> Islamic inscriptions, which are well represented, are excluded from this short survey because they belong to a tradition of their own.

---

Alessandro Bausi, Pietro M. Liuzzo, Universität Hamburg



© 2018 Alessandro Bausi and Pietro M. Liuzzo

This work is licensed under the Creative Commons Attribution 4.0 International license (CC BY 4.0)



and published; notable among them are some new Sabaeen (or Sabaic) inscriptions from Təgrāy (Kropp, 2011), some metal royal inscriptions from the early Aksumite period, and the funerary Greek inscription from Gumālā (Fiaccadori, 2003).

The ancient and medieval inscriptions can be classified as follows:

- 1) 179 items (*RIÉ* nos 1–179) for the pre-Aksumite inscriptions. Most of these inscriptions are in the South Arabian (Sabaeen) language from the first millennium BCE, in the pre-Aksumite period. These inscriptions attest to the presence of Semites (Semitic-speaking people) in the region from the first millennium BCE; moreover, among the Sabaeen inscriptions, a sub-group can be distinguished with linguistic features of its own. In some cases, the same artefact bears both standard Sabaeen and non-standard Sabaeen texts.<sup>2</sup>
- 2) 90 inscriptions (*RIÉ* nos 180–269) from the Aksumite period. They comprise:
  - 2.1) A few early Ethiopic inscriptions (*RIÉ* nos 180–184).
  - 2.2) Also placed in this period are two recently discovered, now published, metal inscriptions (Gebreselassie, 2017; Nebes, 2017). Apparently of great importance, they bear royal names and, along with a previous example that is considered the earliest document of Ethiopic language (*RIÉ* no. 180), they were also inscribed on metal (probably bronze);
  - 2.3) This group includes the great royal inscriptions from Aksum (with only *RIÉ* no. 195 from Marib in Yemen). Of paramount importance for the history of the region (*RIÉ* nos 185–195), they document, in particular, the conversion from a peculiar paganism (at variance with the South Arabian pantheon of pre-Aksumite times) to Christianity of King ‘Ezānā around the first half of the fourth century. They also document the enterprise and military expedition of King Kāleb, especially important for having led to the conquest of South Arabia (Ḥimyar) and to its control by the Aksumites for some years, in the second quarter of the sixth century CE. They also document, with the presence of biblical quotations, the likely accomplishment of the translation of the Bible into Gə‘əz by the early sixth century at the latest. Finally, they document the decay of Aksum with the last larger Aksumite inscriptions, poorly written and at present hardly readable, where linguistic phenomena typical of the later period start to appear. Most of the royal inscriptions were intended to be parts of votive thrones. However, only the evidence of the bases is, to some extent, preserved. The inscriptions, which were probably used as backs and/or side panels of the thrones, were removed over the course of time. They are found at present in various places at Aksum. Some have been discovered as

---

<sup>2</sup> These inscriptions can be found also in the DASI, *Digital Archive for the Study of pre-islamic Arabian inscriptions* database, where 49 inscriptions come from Ethiopia and 26 from Eritrea [<http://dasi.cnr.it/index.php?id=86&prjId=1&corId=0&collId=0&navId=0>]. The map demonstrates the geographical continuity.

they were reused as construction materials in private houses. These inscribed thrones certainly had the function of shaping the landscape and were part of a general plan where the iconic and emblematic meaning of the inscriptions played a particular role. One more point of interest is provided by the likely survival of the introductory protocol, as given by some of these inscriptions, in early medieval documents eventually preserved in Ethiopian archives and of which we have only scanty evidence in additional notes written on blanks of manuscripts, loose leaves and unbound quires. This evidence establishes a suggestive connection between archival practices and inscriptions.

- 2.4) Quite remarkable for its much disputed chronology (dated in a range between the ninth and the fourteenth century CE), the inscription from Ham, in Eritrea (*RIÉ* no. 232) provides an interesting case-study. Once built in the façade of an old half-ruined church dedicated to St Mary, along with other reused materials (including a second Greek inscription containing the monograms A Ω), it has been moved and relocated inside the newly built church at Ham (in 1992). The inscription commemorates the death of a young woman and the selection of biblical passages of the text betrays and presupposes the use of a developed liturgy;
- 2.5) Others, all in Ethiopic (*RIÉ* nos 196–269).
- 3) 17 inscriptions in Greek, from the Hellenistic (one only, *RIÉ* no. 276, *Monumentum Adulitanum*, I) and Aksumite period (*RIÉ* nos 269–286). Among the inscriptions in Greek, the *Monumentum Adulitanum* is particularly remarkable. It consists of two inscriptions, the first is mutilated and the latter is acephalous; the first was issued by Ptolemaeus III (246–222 BCE), who is explicitly mentioned, whereas the latter part is due to an unknown Ethiopian king and was placed upon a throne. Neither of these inscriptions are preserved; they were copied and transmitted by Cosmas Indicopleustes in the *Topographia Christiana*, who also provides information on their material and arrangement. The *Monumentum Adulitanum* is also remarkable for having been used early by Johann Gustav Droysen as exemplary for his definition of “Hellenismus”, in linguistic terms (Canfora, 1995, pp. 15–18). A few years ago, an additional Greek funerary inscription from Gumālā was discovered and published (Fiaccadori, 2003).
- 4) One inscription in a known script (South Arabian of the type used in other Aksumite inscriptions), but in an unknown language (*RIÉ* no. 287), presumably still from the Aksumite period, is known, but not properly deciphered.
- 5) 98 inscriptions on objects (*RIÉ* nos 287–384): 5 seals (*RIÉ* nos 287–291), 18 small, inscribed bronze objects (*RIÉ* nos 292–309), 85 inscriptions on pottery, particularly from the city of Maṣarā in Eritrea (*RIÉ* nos 310–384).
- 6) 59 rock monograms (*RIÉ* nos 385–443), particularly from the region of Qoḥayto in Eritrea.

Excluded from the *RIÉ* repertory are several inscriptions, dated to a later period, on an artefact that is quite peculiar to Ethiopian Christianity, although its models or premises might go back to Coptic Egypt—the *manbara tābot* (plur. *manābərta tābot*), namely “altar chest” (Fritsch, 2010). The oldest altar chests are datable to the twelfth/thirteenth century, in the so-called “Zagwe period”. The *manbara tābot* is like a “chair” (*manbar*) that supports the altar. It can be made from one wooden block, or in rock or metal. The most remarkable examples that bear inscriptions, however, are all in wood. Particularly interesting is the connection between *manbara tābot* inscriptions and parallel texts transmitted in parchment manuscripts.

## 6.2 Beta Maṣāḥəft

The project Beta maṣāḥəft: Manuscripts of Ethiopia and Eritrea (Schriftkultur des christlichen Äthiopiens und Eritreas: eine multimediale Forschungsumgebung)<sup>3</sup> is a long-term project funded within the framework of the Academies’ Program (coordinated by the Union of the German Academies of Sciences and Humanities) under survey of the Akademie der Wissenschaften in Hamburg. The Hiob Ludolf Centre for Ethiopian Studies at the Universität Hamburg hosts the project and aims at creating a virtual research environment that manages complex data related to the predominantly Christian manuscript traditions of the Ethiopian and Eritrean Highlands.<sup>4</sup> The structure of the project is very simple with a TEI encoded XML file for each textual unit, one for each person, place repository and manuscript. Among the records of this last type, there are also inscriptions as they constitute part of the Ethiopian written documentation. The complexity of the corpus of inscriptions related to the scope of this project is evident and no final decision as to the criteria for inclusion and exclusion has been made. The data structure of the project hosts the transcriptions of manuscripts with their descriptions, and the edition of the texts in a separate text edition. This model would fail for inscriptions whose text is much better published directly with the metadata. There are also other projects, like the DASI project, which have already made valuable editions in TEI XML of texts in this corpus, which need to be taken into account in the encoding to guarantee continued interoperability among the existing resources. We will describe in the following section how we plan to encode inscriptions in this context, giving some examples.

<sup>3</sup> [<https://www.betamasaheft.uni-hamburg.de/>].

<sup>4</sup> A preliminary technical description can be found in Liuzzo, 2017.

### 6.3 Inscriptions in Beta Maṣāḥəft

Especially relevant for the project are the inscriptions in ancient Ethiopic language (Gəʿəz), regardless of the script used to write this language. The Greek inscriptions are also included for their historical relevance. The Beta Maṣāḥəft schema already enforces all of the EpiDoc specifications (Elliott et al., 2007) and the editions of texts validating to the project schema are also validated to the latest EpiDoc schema. We will describe here a few of the challenges encountered in the process of including these documents in the framework of the project: 1) the connection to digitally encoded texts that have already faced the problems of encoding a Semitic script (thus the need of working on and encoding the transcription), 2) the encoding of parallel *fidal* and transcribed text and 3) the structuring of the data for the pseudo-trilingual inscription *RIĒ* nos 185, 185bis, 270 and 270bis in the framework of the current project.

#### 6.3.1 The Challenges of Encoding Inscriptions in Semitic Scripts

Inscriptions in Sabaean from Eritrea and Ethiopia are already published online within the DASI project (Avanzini et al., 2014). The Beta Maṣāḥəft project does not currently include those texts directly, but will include links to the DASI editions online by means of a simple *<ref>* element in a host XML file with the references. This has been produced from the XML corpus export by the project from which only the local ID, the main reference, the *<respStmt>* and titles were taken to make a mini record with a link to the actual resource in the DASI project website. In fact, although both projects work in TEI XML, the terms of use of the data and the structure of the records does not allow for a direct import of the data.<sup>5</sup>

The latter project has developed a highly sophisticated encoding method for the onomastic features, which is perfectly consistent with the mark-up practices of Beta Maṣāḥəft and validates to its schema, although the scope of the project does not currently allow for as deep an annotation as the one carried out in DASI. The inscriptions in Gəʿəz encoded in Beta Maṣāḥəft follow this encoding structure, especially for the techniques identified for overlapping semantic mark-up, allowing for a cross analysis of the inscriptions in the two projects.

The mark-up needs to be carried out on the transcription for these texts, especially with regard to the morphological aspects; this is also the approach taken by the TraCES project<sup>6</sup> (Bausi, 2015) which includes morphologically annotated texts

---

<sup>5</sup> Further cooperation is envisaged to integrate the collections and the mark-up of the inscriptions in the two projects.

<sup>6</sup> [<https://www.traces.uni-hamburg.de/>] founded by the European Commission ERC-AG-SH5 - ERC Advanced Grant - Cultures and cultural production, grant number 338756.

of inscriptions elaborated by Maria Bulakh. Once these last annotations are imported into Beta Maṣāḥəft it will be possible to interrogate the onomastic features annotated in DASI, together with the morphological features.

Although it is, in principle, no problem to annotate the transliteration instead of the text in the original script, the current search functionalities of the Beta Maṣāḥəft online application<sup>7</sup> prefer the *fidal* script and cannot perform a bidirectional conversion between *fidal* and transliteration for search purposes.<sup>8</sup> To guarantee the presence of both an annotated text in *fidal* and transliteration (both are needed for the aims of our project and for interoperability purposes described above), texts of inscriptions are reproduced in both scripts.<sup>9</sup> The following is an example with the first three lines of the inscription *RIÉ* 187

```
<ab xml:lang="gez-trsl">
  <lb n="1"/> <persName ref="PRS3938Ezana"><supplied reason="lost">' ezānā walda
  <persName ref="PRS3729ellaAm">'ēle 'amidā</persName> bə' əsayā ḥalen nəḡuša
  <placeName ref="LOC1310Aksum">'ak<lb
  n="2" break="no"/>sum</placeName> waza <placeName ref="LOC3868Himyar">ḥəmer</placeName>
  waza <placeName ref="LOC5333Raydan">raydān</placeName> waza <placeName
  ref="LOC5395Saba">saba </placeName>
  waza <placeName ref="LOC5491Salhen">salhen</placeName> wa<lb
  n="3" break="no"/>za <placeName ref="ETH2065seyamo">šəyāmo</placeName></supplied> waza
  <placeName ref="ETH2263bega">bəḡā</placeName> waza <placeName ref="ETH1768Kasu-
  K">kāsu</placeName> <roleName type="title">nəḡuša
  <supplied xml:id="sup1" next="sup2" reason="lost">nagašt</supplied></roleName></persName>
  <supplied xml:id="sup2" prev="sup1" reason="lost">wa</supplied>
</ab>

<ab xml:lang="gez">
  <lb n="1"/> <persName ref="PRS3938Ezana"><supplied reason="lost">ዳኒአን ወልደ
  <persName ref="PRS3729ellaAm">ዳኒአን ወልደ</persName> ቤላሳላ ስላሳ ስላሳ ገሥት <placeName
  ref="LOC1310Aksum">አክ<lb
  n="2" break="no"/>ሙር</placeName> ወዘ፡ <placeName ref="LOC3868Himyar">ሕማር</placeName> ወዘ፡
  <placeName ref="LOC5333Raydan">ረይዳን</placeName> ወዘ፡ <placeName ref="LOC5395Saba">ሰበሌ፡
  </placeName> ወዘ፡
  <placeName ref="LOC5491Salhen">ሰልኬን</placeName> ወ<lb
  n="3" break="no"/>ዘ፡ <placeName ref="ETH2065seyamo">ሻይሞ</placeName></supplied> ወዘ፡
  <placeName ref="ETH2263bega">ቤገ፡</placeName> ወዘ፡
  <placeName ref="ETH1768Kasu-K">ካሱ፡</placeName> <roleName type="title">ገሥት፡ <supplied
  xml:id="sup1g" next="sup2g" reason="lost">ነገሥት፡</supplied></roleName></persName>
</ab>
```

The parallel mark-up shows the identified named entities and provides data that can be queried to list forms of the title of the king, for example.

<sup>7</sup> Not yet available online. Data is available, with full documentation, here: [https://github.com/BetaMasaheft].

<sup>8</sup> This task is currently being elaborated under the TraCES project.

<sup>9</sup> The transliteration is produced with code available also via this self-standing application: [https://betamasaheft.github.io/transliteration/].

### 6.3.2 Multilingual Inscriptions

We will look now at the example of *RIÉ* 185, 185bis, 270 and 270bis. The texts of *RIÉ* 185 and *RIÉ* 270 and *RIÉ* 185bis and 270bis respectively, have been grouped in a single record for each stone to follow the praxis of one record for each manuscript. The concordance to the original references is preserved in the data. The internal text structure has been maintained inside the record, as also in *RIÉ*, instead of duplicating the record for the scripts employed. There are then two records in TEI XML for these two stones, which contain three parts each, and represent the actual distribution of the text on the different faces of the stone support.

The first problem posed by these texts is the relation between the three copies of the text in different scripts and languages and the relation between the main copy and the second copy. The first problem, as well as the changing text direction, is encoded in the diplomatic edition using *@xml:lang*. The second aspect is encoded in the XML data of Beta Maṣāḥəft using the relation element and properties from the SAWS ontology<sup>10</sup> in its *@name* attribute.

```
<relation name="saws:isDirectCopyOf" active="RIE185bisand270bis" passive="RIE185and270"/>
```

The text has been edited in Ethiopic and Greek, and our record reflects this, leaving the transcription of the texts in their diplomatic form in the record about the stones, and the edition in a text. The relation among the stones and the text is made by means of the *<listWit>* element in the work records, and using *@corresp* attributes in the inscriptions records, corresponding to each relevant text part, as is generally the practice for the manuscripts and their contents' annotation in the project. The Gə'əz text will thus contain a list of witnesses.

```
<listWit>
  <witness xml:id="A" corresp="RIE185and270#RIE185!"/>
  <witness xml:id="B" corresp="RIE185and270#RIE185!!"/>
  <witness xml:id="C" corresp="RIE185bisand270bis#RIE185bis!"/>
  <witness xml:id="D" corresp="RIE185bisand270bis#RIE185bis!1 RIE185bisand270bis#RIE185bis!2
RIE185bisand270bis#RIE185bis!3"/>
</listWit>
```

Here, it may be observed that we have to provide several IDs in the *@corresp* attribute of the witness D, because in the description of this document, the actual text of this version is split over three faces of the stone. The structure of the text on the stone is reflected in the description of the inscription in XML as follows (text has been omitted).

<sup>10</sup> [http://www.ancientwisdoms.ac.uk/].

```

<div type="textpart" subtype="face" xml:id="A">
  <div type="textpart" n="1" xml:id="RIE270bis" xml:lang="grc" corresp="LIT4851greekRoyal">
    <ab></ab>
  </div>
  <div type="textpart" n="5" xml:id="RIE185bisII3" xml:lang="gez" corresp="LIT4850pseudotrilingual">
    <head>D part 3</head>
    <ab></ab>
  </div>
</div>

<div type="textpart" subtype="face" xml:id="B">
  <div type="textpart" n="2" xml:id="RIE185bisI" xml:lang="gez-sabaic" corresp="LIT4850pseudotrilingual">
    <head>C</head>
    <ab></ab>
  </div>

  <div type="textpart" n="3" xml:id="RIE185bisII1" xml:lang="gez" corresp="LIT4850pseudotrilingual">
    <head>D part 1</head>
    <ab></ab>
  </div>
</div>

<div type="textpart" subtype="face" xml:id="C">
  <div type="textpart" n="4" xml:id="RIE185bisII2" xml:lang="gez" corresp="LIT4850pseudotrilingual">
    <head>D part 2</head>
    <ab></ab>
  </div>
</div>

```

Note that the letters indicated traditionally for the four texts, are preserved both as abbreviation for the apparatus and as headers in the diplomatic edition. Also, the denomination of the different faces of the stone are preserved as *@xml:id* of the relevant text part.

The relation between the two texts, the Gəʾəz and the Greek text, is stated in both records by means of another relation element.

```

<relation name="saws:isVersionInAnotherLanguageOf" active="LIT4850pseudotrilingual"
  passive="LIT4851greekRoyal"/>

```

The editions of the texts do not need a line division, neither do they need to follow one of the scripts used, but the XML file can host any combination. The existing edition of the Gəʾəz text elaborated by A. Bausi uses the transliteration as the *RIĖ* texts and it is encoded in this way.

### 6.3.3 Inscriptions in Greek

Further issues are presented by the encoding of Greek texts like *RIĖ* 276, mentioned above, known only through manuscript tradition. In this case, we may have only a work record in Beta Maṣāḥəft with the text of an edition of the inscription. This allows the text to be linked to other resources and reflects its status without forcing the presence of an inscribed support. However the manuscript does provide this information and we encode them in a specific manuscript record as in the previous example. The encoding of these texts does not present special issues and follows the schema of the project validating also to EpiDoc.

## 6.4 Conclusions

Representing inscriptions along with manuscripts in The Beta Maṣāḥəft project using TEI, while posing challenges, provides a clear and documented XML representation of the information, enabling connection with other XML resources like those in the DASI project. Ahead of us lies the challenge of integrating the encoding used for inscriptions in the DASI project with the morphological annotation exported from the TraCES project inside the Beta Maṣāḥəft project data structure and schema.

## Bibliography

- Avanzini, A., De Santis, A., Marotta, D., & Rossi, I. (2014). Between Harmonization and Peculiarities of Scientific Domains. Digitizing the Epigraphic Heritage of pre-Islamic Arabia in the Project DASI. In S. Orlandi, R. Santucci, V. Casarosa, & P.M. Liuzzo (Eds.), *Information Technologies for Epigraphy and Cultural Heritage: Proceedings of the First EAGLE International Conference* (Serie antichistica. Collana Convegni 26) (pp. 69–93). Roma: Sapienza Università Editrice. Retrieved from [<https://www.eagle-network.eu/wp-content/uploads/2015/01/Paris-Conference-Proceedings.pdf>], 2017/11/30. doi: 10.13133/978-88-98533-42-8
- Bausi, A. (2015). TraCES: From Translation to Creation: Changes in Ethiopic Style and Lexicon from Late Antiquity to the Middle Ages. In A. Bausi, A. Gori, D. Noshitsin, & E. Sokolinski (Eds.), *Essays in Ethiopian Manuscript Studies. Proceedings of the International Conference Manuscripts and Texts, Languages and Contexts: the Transmission of Knowledge in the Horn of Africa, Hamburg, 17-19 July 2014* (pp. 11–13). Wiesbaden: Harrassowitz Verlag.
- Bernand, É., Drewes, A. J., & Schneider, R. (1991). *Recueil des Inscriptions de l'Éthiopie des périodes pré-axoumite et axoumite*. Introduction de Fr. Anfray; I: Les documents; II: Les Planches; Étienne Bernand, id., III: Traductions et commentaires, A: Les inscriptions grecques. Paris: Diffusion de Boccard.
- Canfora, L. (1995). *Ellenismo*. Bari: Laterza.
- Elliott, T., Bodard, G., Milonas, E., Stoyanova, S., Tupman, C., & Vanderbilt, S. (2007, 2013). *EpiDoc Guidelines: Ancient documents in TEI XML*. Retrieved from [<http://www.stoa.org/epidoc/gl/latest/>], 2017/12/09.
- Fiaccadori, G. (2003). Un'epigrafe greca aksumita (RIÉth 274). In V. Ruggeri & L. Pieralli (Eds.), *Eukosmia. Studi miscellanee per il 75° di Vincenzo Poggi S.J.* (pp. 243–255). Soveria Mannelli: Rubbettino Editore.
- Fritsch, E. (2010). Tabot: Mänbärä tabot. In A. Bausi & S. Uhlig (Eds.), *Encyclopaedia Aethiopica* (Vol. 4: O-X, pp. 804b–807a). Wiesbaden: Harrassowitz Verlag.
- Gebreselassie, Y. (2017). L'alphabet éthiopien. Une origine discutée. *Dossiers d'Archéologie*, 379 = *Éthiopie: un patrimoine exceptionnel*, 34–37.
- Kropp, M. (2011). Schriften und Sprachen im Kontakt: Sabäisch in Äthiopien und die ersten Zeugnisse der äthiopischen Sprache und Schrift. In S. Wenig (Ed.), *In kaiserlichem Auftrag. Die Deutsche Aksum-Expedition 1906 unter Enno Littmann* (Vol. II, pp. 323–337). Wiesbaden: Reichert Verlag.
- Liuzzo, P.M. (2017). Encoding the Ethiopic Manuscript Tradition. In *Proceedings of Balisage: The Markup Conference 2017* (Balisage Series on Markup Technologies 19). doi: 10.4242/BalisageVol19.Liuzzo01
- Nebes, N. (2017). The Inscriptions of the Aksumite King Ḥafil and their Reference to Ethio-Sabaeen Sources. *Zeitschrift für Orient-Archäologie*, 10, 356–369.



Paolo Xella and José Á. Zamora

## 7 Phoenician Digital Epigraphy: CIP Project, the State of the Art

**Abstract:** The corpus of Phoenician-Punic inscriptions comprises about 12,000 documents, spread over a very wide area and span of time (all the countries of the Mediterranean region, from the end of the 2<sup>nd</sup> millennium BCE to the first centuries of the 1<sup>st</sup> millennium CE). The quantity and nature of the documents have caused considerable difficulties in the knowledge and scientific use of these sources. The project CIP (*Corpus Inscriptionum Phoenicarum necnon Poenicarum*, also known as the PhDB or Phoenician Data Base) came into being to tackle these problems by producing a collection and a critical edition of all the epigraphic documents in the form of a data bank.

**Keywords:** epigraphy, North-West Semitics, Phoenician & Punic, corpus, data bank

### 7.1 Motive of the Project and Institutional Background

According to generally accepted estimates, the corpus of Phoenician-Punic inscriptions comprises about 12,000 inscriptions from all the countries of the Mediterranean. As noted when this project was presented (Cunchillos, Xella, & Zamora, 2005), the sheer quantity and scattered nature of the documents, spread over a very wide span of time, have severely affected research and caused considerable difficulties in the knowledge, availability and use of these sources. In fact, as yet there is not even a simple, complete and reliable list of existing Phoenician inscriptions, still less a critical edition of them. There are only incomplete collections or anthologies (Cunchillos, Xella, & Zamora, 2005, pp. 517-519, with references) most of which need bringing up to date.

This lack of verified documents has had repercussions on the very knowledge of the Phoenician language, making it extremely difficult – and in some cases even impossible – to revise and update basic study tools (such as grammars, dictionaries, concordances, etc.). This state of affairs has seriously restricted the role that epigraphic evidence – the only direct written source of Phoenician culture – should play in general historical information.

In order to tackle these problems and to try to resolve them, by making Phoenician texts – presented with rigorous and uniform criteria – available to the academic

---

**Paolo Xella**, Consiglio Nazionale delle Ricerche; Universität Tübingen

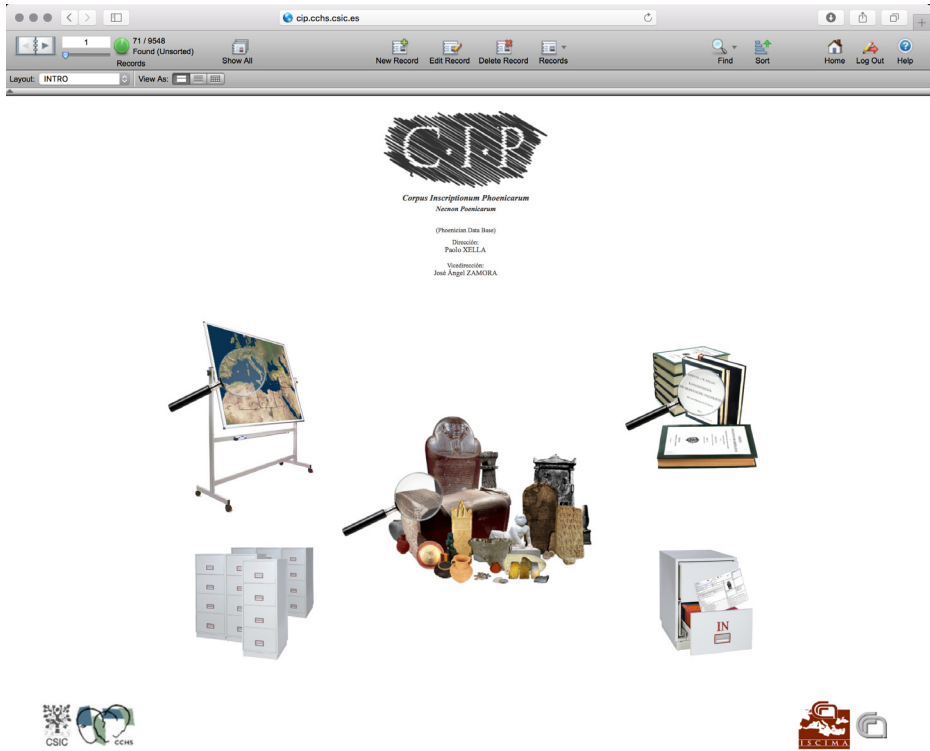
**José Á. Zamora**, Escuela Española de Historia y Arqueología en Roma; Consejo Superior de Investigaciones Científicas



© 2018 Paolo Xella and José Á. Zamora

This work is licensed under the Creative Commons Attribution 4.0 International license (CC BY 4.0)

community, the project CIP – *Corpus Inscriptionum Phoenicarum necnon Poenicarum*, also known as the PhDB or *Phoenician Data Base*, came into being (Figure 7.1).<sup>1</sup> It was born from an Italian initiative that early on made use of collaboration with a CSIC Spanish team (Cunchillos, Xella, & Zamora, 2005; Xella & Zamora, 2007).



**Figure 7.1:** Home menu of the CIP data bank, with basic access links

## 7.2 Aims and General Description of the Project

The project produces a collection and a critical edition of all Phoenician and Punic epigraphic documents in the form of a data bank (realistically speaking, the only form possible). The data bank also aims to include all available information on every Phoenician epigraphic document, presenting the relevant data in an ordered and programmatic form, together with graphic and photographic material.

<sup>1</sup> [<http://cip.cchs.csic.es/>].

In this electronic edition, the inscriptions are collected and classified according to a uniform and open-ended system, which allows updating of the various regional corpora in real time. This can be done either by inserting new documents or by checking and expanding available information, improving known readings thanks to collations, and by extending data or bibliography in various ways. Each epigraphic document is identified by a unique code, based on a single criterion, with cross-reference to previous editions or collections, together with the most complete and up-to-date bibliography. All the information about the inscriptions can be requested and analysed from various aspects and on different levels (Figure 7.2).

The aim of the CIP is to present a text based on original collations. Whenever it is impossible to collate a document directly, the CIP proposes a textual version resulting from other ways of checking (based on photographs, engravings, copies, etc.), or else based on one or more editions that are expressly indicated in a specific field. As a result, the text presented in the Database is a genuine edition of it, involving the use of a whole set of conventions and critical choices. These parameters are based on those normally used by specialists, with minimal adaptations to the demands of an electronic format. This is in order to accommodate further computerized analysis, such as the generation of automatic segmentation, restorations, concordances, morphological analysis, etc. From a technical point of view, it must be noted that all the textual information is stored and managed in linked “tables” (taking advantage of the use of a relational database).

### 7.3 Basic Technical Data

Such a difficult and wide-ranging project needs the application of new technology (with a precise and suitable methodology, cf. for example Cunchillos, 2000).

Technically speaking and without going into detail, it should be noted that the CIP project uses customized applications for processing data. All the data banks created rely upon well-known commercial software: relational databases with client-server architecture, programmed *ad hoc*. In this way, it is possible to generate new data from the data already entered, to organize the records functionally (in linked “tables”) and to share them, having available a single block of data brought up-to-date in real time. Members of the team can access this main block of data on-line thanks to a simple web browser, independently of the type of hardware or system-software used (Zamora, 1997, 2007).



## 7.4 Organization and Structure of the Corpus

In the last few years, the decision to use various kinds of information technology has prevented the enormous amount of material accumulated during the various phases of the project from generating dramatic problems in managing the documentation. Instead, it has turned into a genuine “open-ended catalogue”. It is a sort of working edition that has proved to be very useful, right from the start, not only for the development of the project itself, but also, in general, for epigraphic, linguistic and historical research.

As noted above, the organization of the material has been achieved by adapting the tools of information technology to the methodology and aims of epigraphy, as the very structure of the corpus demonstrates. Each document has a main file (a “record”) in a primary table, which contains a set of information arranged (in “fields”) and standardized and normalized where possible and useful: this basic information concerns both the inscribed object (from the date and place where it was found to its formal and material characteristics) and the text (from its transcription to the relevant bibliography), and is linked with more information arranged and distributed in other files (organized in separate related tables). Different layouts allow the user to integrate all the available information in several practical ways (Figure 7.3).

The normalization of part of the information is by no means arbitrary. By distinguishing and standardizing specific data, we have tried to set up reliable and effective criteria for selecting the documents. In this way, uniform research on groups of inscriptions has been possible based, for example: on their find-spot or date, type of inscribed object, technique, material, etc. Research can be carried out on a single topic or on a combination of topics. Research on any other type of information (material, bibliography, text, etc.) or by selecting other criteria is always possible as well.

The dominant criterion for organising these documents is based on the find-spot (conventionally called “locality”) of every inscribed object. As for other data, a linked table has been created (so allowing it to be managed separately) with records for each locality containing more information (together with photographs and maps). This find-spot is part of a geographical area (field: “nation” or “country”) corresponding to the modern country to which it belongs. Although this criterion is only remotely based on historical and cultural reasons (and thus just some meaningful areas of the past vaguely coincide with nations of the present), it allows for a simple and objective classification. Within each nation, when considered useful or necessary, a further regional subdivision is used (field: “region”) identified by a number, which is not based on historical criteria but on clear geographical and/or administrative criteria (Figure 7.4).

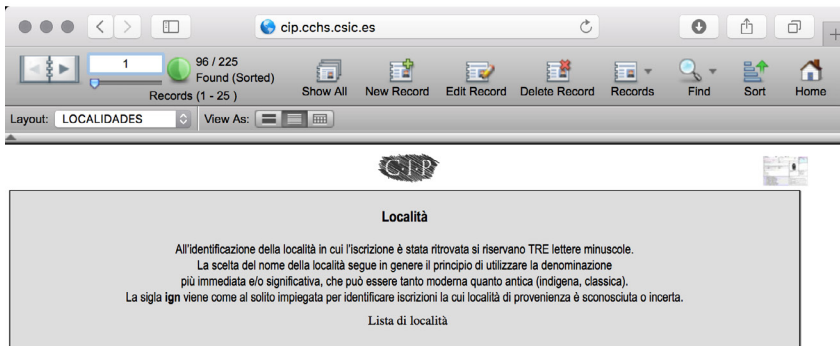
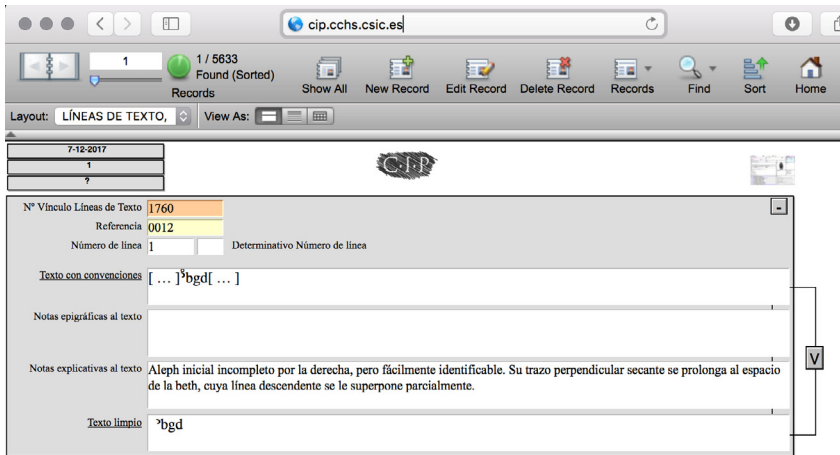
The screenshot shows the CIP Project web interface. At the top, there is a navigation bar with options like 'New Record', 'Edit Record', 'Delete Record', and 'Records'. Below this, there is a search bar and a 'Layout: GENERAL' dropdown. The main content area is divided into several sections:

- Header:** 'es 2 tdb 0012' and 'Aut. / Fac. / Pat. / Photo. / Entry'.
- Location:** 'Lugar: Parte suroccidental de la ciudad, en la vertiente norte del espigón del puerto (Esp.1.V/Norte)'. 'Data: 1986 (foto)'. 'Cód. TDB: 86001'.
- Material:** 'Material: Cerámica / ceramic / ceramic / céramique'. 'Tip: Recipiente abierto / recipiente abierto / open container / récipient ouvert'. 'Esp: 54'. 'Luz: 47'. 'Luz: 47'.
- Description:** 'Lugar: Vertiente norte del espigón del puerto, en un estrato de sillón de material bastante uniforme que cubren las casas fenicias. Se trata de un paquete de tierra ocre con escasa cal, bastante homogéneo. En el estrato apareció mucho material y escasa piedra. Esta limitado por el muro L al norte y por el muro G al sur.'.
- Date:** 'Data: -725 (Finis del siglo VIII o principios del VII a. C.)'. 'Data: -675 (probablemente principios del VII a. C.) por (probablemente principios del VII a. C.) por'.
- Bibliography:** 'Bibliografía: Cunchillos, 1991: 13-22 (+EP). Cunchillos, 1994: 207, 209, 212, 213, 215, 216. Cunchillos - Vila, 1998: 31-38 (como ejemplo). Cunchillos - Zamora, 2004: 111-114 (data arqueológica, contextualización), 117 (data 17), 120 (ova 20); 122 (ova 49); 124 (Zamora, 2006): 151-164 (contextualización, valor histórico); 175 (texto y ova 41); 176, 177, fig. 5; 178; 182. Zamora, 2011 (top: bibliografía)'. 'Cunchillos, 1991: 14 (foto), 15 (+O. Albaladejo)'. 'Cunchillos - Zamora, 2004: 111, fig. 5. Situación: FORTALEZA'. 'Zamora, 2006: 177, fig. 5.'.
- Notes:** 'Nota: Lectura CIP (JLCO/AZ)'. 'Traducción interpretación (JAZ): El contexto, soporte y parámetros epigráficos del área hacen pensar en una inscripción de propiedad. La sucesión de signos legibles tendría entonces que identificarse con un nombre fenicio (del tipo Abiqad) al que pudo incluso antecederse una preposición de pertenencia ʾ, perdida al fragmentarse el cuenco. Sin embargo, no hay testimonio claro de tal anteposición y su mera existencia teórica es problemática. Por ello, debe considerarse como una opción muy probable que se trate de un abecedario o alfabetario, dada la fácil lectura de la secuencia aliph, beth, gimmet, dalet (que podría aparecer en otros epígrafes de Doña Blanca, cf. tdb ).'.

Figure 7.3: Inscription main layout, integrating data from various primary and secondary tables

This threefold classification (“nation”, “region”, “locality”), completed by a progressively increasing number, forms the set of initials and numbers, or coding sequence, tagging every single inscription, which thus receives a (unique) alphanumeric code that identifies it, producing a “code” or “siglum” of the type lb1byb0001 (= Lebanon/region number 1/Byblos/inscription number 1). In addition, the Database envisages the insertion of cross-references to other collections (for example: CIS I 158 = ICO Sard. 24 = KAI 67 = KI 62). For this purpose, another linked secondary table has been created. The inventory number of the inscribed object in a museum or excavation is also entered into the appropriate field (to facilitate working with groups of inscriptions kept in the same place, for example) linked to another secondary table. Generally speaking, all the standardized fields (material on, and technique by which the inscription is written, type of inscribed object, etc.) are linked to a separate group of files, arranged in a secondary table, with more information.

Even though the CIP does not claim to become a database of epigraphic images (for which there are other projects), it also allows for the insertion of photographs, drawings and various other graphic materials intended especially for rapid identification of the inscribed object (Figure 7.5). In a similar way, the CIP takes advantage of including all the bibliography judged to be relevant for each document to construct a separate (but always linked) bibliographical table.



Località	Sigla	Nazione	Regione			
Bajo de la Campana	bca	Spagna / España / Spain / Espagne	Penisola / Peninsula / Péninsule	2	5	+
Baza	baz	Spagna / España / Spain / Espagne	Penisola / Peninsula / Péninsule	2	1	+
Binatram	bnt	Spagna / España / Spain / Espagne	Isola / Islas / Islands / Îles (es)	1	1	+
Binicodrell Nou	bnc	Spagna / España / Spain / Espagne	Isola / Islas / Islands / Îles (es)	1	1	+
Biniparratx Petit	brp	Spagna / España / Spain / Espagne	Isola / Islas / Islands / Îles (es)	1	5	+
Binissafüller	bln	Spagna / España / Spain / Espagne	Isola / Islas / Islands / Îles (es)	1	22	+
Boades	boa	Spagna / España / Spain / Espagne	Penisola / Peninsula / Péninsule	2	1	+
Bobadilla, La	bob	Spagna / España / Spain / Espagne	Penisola / Peninsula / Péninsule	2	4	+
Ca Na Rafala	cnr	Spagna / España / Spain / Espagne	Isola / Islas / Islands / Îles (es)	1	1	+
Cádiz	cad	Spagna / España / Spain / Espagne	Penisola / Peninsula / Péninsule	2	27	+
Cales Coves	cco	Spagna / España / Spain / Espagne	Isola / Islas / Islands / Îles (es)	1	1	+
Camas, Las	cam	Spagna / España / Spain / Espagne	Penisola / Peninsula / Péninsule	2	1	+
Camposoto	cmp	Spagna / España / Spain / Espagne	Penisola / Peninsula / Péninsule	2	1	+
Can Berri d'en Sergent	cbs	Spagna / España / Spain / Espagne	Isola / Islas / Islands / Îles (es)	1	1	+
Can Fita	cfi	Spagna / España / Spain / Espagne	Isola / Islas / Islands / Îles (es)	1	1	+
Cañares, Los	can	Spagna / España / Spain / Espagne	Penisola / Peninsula / Péninsule	2	1	+
Cancho Roano	can	Spagna / España / Spain / Espagne	Penisola / Peninsula / Péninsule	2	2	+
Cap Negret	crng	Spagna / España / Spain / Espagne	Isola / Islas / Islands / Îles (es)	1	1	+
Carambolo, El	crb	Spagna / España / Spain / Espagne	Penisola / Peninsula / Péninsule	2	7	+

Figure 7.4: Top: Line of text file (Basic layout). Below: List of find-spots (example of layout integrating data from secondary tables)

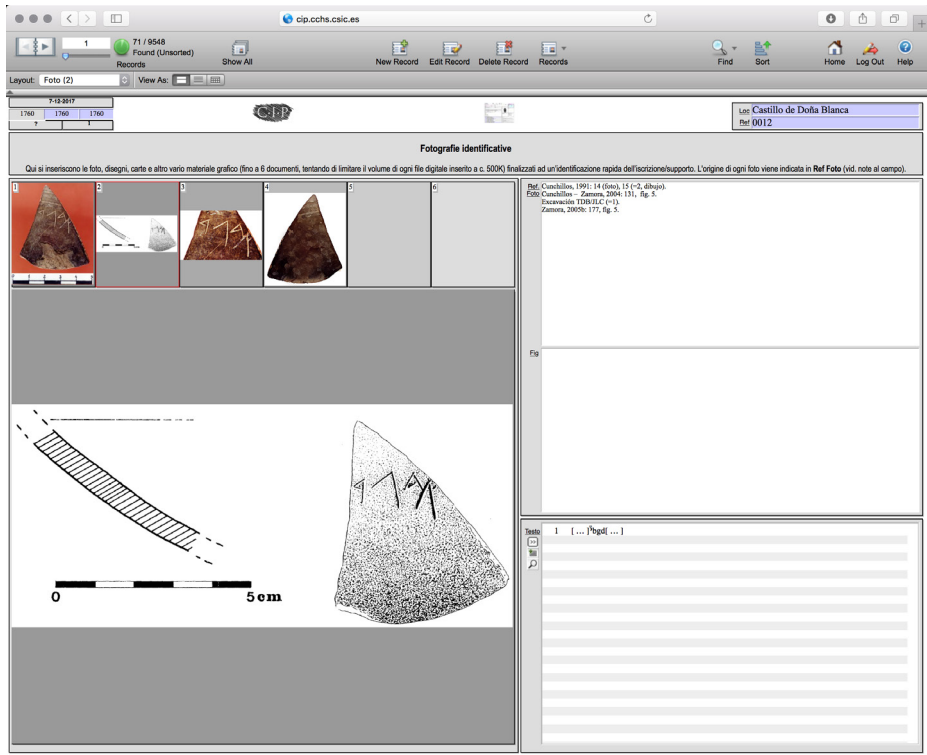


Figure 7.5: Graphic material main layout

## 7.5 State of the Database and Future Outlook

So far, the CIP has collected a total of more than 9,500 inscriptions, and some regional corpora have been catalogued almost completely. However, there are various levels of work: the corpora with fewer documents have made it easier to produce satisfactory critical editions, as is also the case of corpora that have benefited from research projects with epigraphic implications. Where the mass of documents and the resulting difficulties of collection are greater, instead, the project has indicated setting up a base of information to be improved critically at a later phase.

The future outcome of the project includes not only on-line consultation of the whole corpus but, in addition, conventional and electronic publication of catalogues, regional corpora and other research tools extracted from the Database or derived from the project that will be considered useful for research purposes. In the mainframe of the project, a collective monograph in two volumes, aiming to offer a wide overview of current knowledge on Phoenician epigraphy, is about to be published (Amadasi, Xella & Zamora, forthcoming).



## Bibliography

- Amadasi, M.G., Xella, P., & Zamora, J.Á. (Eds.). (forthcoming), *Phoenician Epigraphy. Current knowledge on Phoenician epigraphic evidence in the frame of the project Corpus Inscriptionum Poenicarum necnon Poenicarum. Studi Epigrafici e Linguistici*, 35–36.
- Cunchillos, J.-L. (2000). *Hermeneumática*. Madrid: Consejo Superior de Investigaciones Científicas.
- Cunchillos, J.-L., Xella, P., & Zamora, J.Á. (2005). Il *corpus* informatizzato delle iscrizioni fenicie e puniche: un progetto italo-spagnolo. In A. Spanò Giammellaro (Ed.), *Atti del V Congresso Internazionale di Studi Fenici e Punic (Marsala – Palermo, 2-8 ottobre 2000)* (pp. 517–521). Palermo: Università di Palermo.
- Xella, P., & Zamora, J.Á. (2007). The Phoenician Data Bank: The International Project *Corpus Inscriptionum Poenicarum necnon Poenicarum. Ugarit-Forschungen*, 39, 773–790.
- Zamora, J.Á. (1997). Banco de Datos Filológicos Semíticos Noroccidentales: Fenicio. Primeros módulos del software *Melqart*. In J.-L. Cunchillos, J.M. Galán, & J.Á. Zamora (Eds.), *El Mediterráneo en la antigüedad: Oriente y Occidente (Actas del I Congreso Español de Antiguo Oriente Próximo, Madrid 29 de septiembre - 2 de Octubre 1997)*. Madrid: Centro de Estudios del Próximo Oriente [CD-ROM]. Retrieved from [[https://www.researchgate.net/publication/39352657\\_Banco\\_de\\_Datos\\_Filologicos\\_Semiticos\\_Noroccidentales\\_Fenicio\\_Primeros\\_modulos\\_del\\_software\\_Melqart](https://www.researchgate.net/publication/39352657_Banco_de_Datos_Filologicos_Semiticos_Noroccidentales_Fenicio_Primeros_modulos_del_software_Melqart)], 2018/02/16.
- Zamora, J.Á. (2007). Algunas notas técnicas sobre el *Corpus Inscriptionum Poenicarum necnon Poenicarum (CIP) / Phoenician Data Base (PhDB)*. In J.J. Justel, B.E. Solans, J.P. Vita, & J.Á. Zamora (Eds.), *Las aguas primigenias: El próximo Oriente antiguo como fuente de civilización (Actas del IV Congreso Español de Antiguo Oriente Próximo)* (pp. 203–217). Zaragoza: Instituto de Estudios Islámicos y del Oriente Próximo. Retrieved from [<http://digital.csic.es/bitstream/10261/9226/1/Algunas%20notas%20t%C3%A9cnicas%20sobre%20el%20Corpus%20Inscriptionum%20Poenicarum%20necnon%20Poenicarum%20%28CIP%29.pdf>], 2018/02/16.

Daniel Burt, Ahmad Al-Jallad and Michael C.A. Macdonald

## 8 The Online Corpus of the Inscriptions of Ancient North Arabia

**Abstract:** The *Online Corpus of the Inscriptions of Ancient North Arabia* (OCIANA) was created to make available in a fully searchable online database the texts and translations of all the inscriptions of ancient North Arabia, together with metadata and photographs. Developed in Filemaker Pro, it is consultable both online and as a series of fully searchable pdfs. All known inscriptions from ancient North Arabia have been entered, except the “Thamudic”, which pose particular problems, and will be entered in the next phase of the project.

**Keywords:** Ancient North Arabian, online database, lexicography, ancient literacy, glyph variation

This chapter is in three parts. In the first, Michael Macdonald describes the origins and purpose of the *Online Corpus of the Inscriptions of Ancient North Arabia* (OCIANA).<sup>1</sup> In the second, Daniel Burt describes its present structure and performance, and in the third Ahmad Al-Jallad looks forward to the aims of Phase 3 of the project, which he will direct.<sup>2</sup>

### 8.1 The Background to OCIANA

OCIANA aims to make available in one place an edition of all known inscriptions from ancient North Arabia. The term “ancient North Arabia” in this context refers geographically to the Arabian Peninsula north of Yemen,<sup>3</sup> with a fluid northern

---

1 [<http://krc2.orient.ox.ac.uk/ociana/>].

2 Phase I was a preparatory stage lasting one year (2011–2012) and funded by the John Fell Fund of the University of Oxford. During this phase the *Ancient Arabia: Languages and Cultures* (AALC, <http://www.ancientarabia.co.uk/>) website was established and approximately 10,000 black-and-white negatives and colour slides of previously unpublished Safaitic inscriptions from the Basalt Desert Rescue Survey (BDRS) were scanned in preparation for their insertion in OCIANA during Phase 2.

3 Ancient South Arabian inscriptions have been collected and edited in the *Digital Archive for the Study of pre-Islamic Arabian Inscriptions* [<http://dasi.cnr.it/>], based at the University of Pisa and *The Sabaic Dictionary Online* which are described in Chapters 1 and 9 in this book.

---

Daniel Burt, Michael C.A. Macdonald, University of Oxford  
Ahmad Al-Jallad, Universiteit Leiden



border including modern Jordan, southern Syria and western Iraq. Chronologically, it refers to inscriptions in all languages and scripts from this area before the Islamic era.

This means that while the vast majority of the inscriptions in OCIANA are, and will continue to be, in the Ancient North Arabian [ANA] scripts (see below), it will also contain those texts in Akkadian, Old Aramaic, Imperial Aramaic, local forms of the Aramaic script, Nabataean, Palmyrene, Greek, and Latin that have been found in Arabia, north of Yemen.

The ANA scripts are varieties of the “South Semitic script-family”, which separated from the North-West Semitic (Phoenico-Aramaic) branch shortly after the invention of the alphabet, and developed in parallel to it. In antiquity, it was used solely in Arabia and its immediate surroundings, and its only modern survivor is the vocalized alphabet used in Ethiopia for Gəʿəz, Amharic and other languages (Macdonald, 2008). In antiquity, one form of the South Semitic script-family was used in southern Arabia – the *musnad*, or Ancient South Arabian [ASA] “monumental” script, from at least the tenth century BCE (Stein, 2013). From this then developed the *zabūr*, a form of the script used to carve everyday documents on the stems of palm-leaves or on sticks (Stein, 2005a, 2005b). In the east of Arabia, between the Saudi Arabian oasis of al-Ḥasā and the Oman Peninsula, the ASA script was used to express what may be a North Arabian language, “Hasaitic”, alongside Aramaic (Overlaet, Macdonald, & Stein, 2016, pp. 132–140).

However, in the western two-thirds of Arabia, north of Yemen, a number of different alphabets developed from the South Semitic script-family and these were used by the inhabitants of oases in north-west Arabia (Dadan – modern al-ʿUlā; Figures 8.1 and 8.2 – Taymāʿ, and probably Dūmah – modern Dūmat al-Jandal/al-Jawf).



**Figure 8.1:** Dadanic inscriptions at al-ʿUdhayb (al-ʿUlā, Saudi Arabia). (Photograph by C.J. Robin)



**Figure 8.2:** Detail of Fig. 8.1, Dadanitic inscriptions at al-‘Udhayb (al‘Ulā, Saudi Arabia). (Inscriptions U 011–019, 021–026, see OCIANA). (Photograph by C.J. Robin)

A number of different scripts of the same family were also used very widely among the nomads from southern Syria to Yemen, who were literate at different times in different areas during the second half of the first millennium BCE and the fourth century CE.<sup>4</sup> By far the most numerous of these graffiti by nomads are the “Safaitic inscriptions” (Figures 8.3 and 8.4), which are found in their tens of thousands on the rocks of the deserts in southern Syria, north-eastern Jordan, and northern Saudi Arabia (Macdonald, 2010).

Almost certainly, more people in North Arabia were literate during this period than in any other part of the Middle East, and they have left us vast numbers of inscriptions. Yet, despite this, the history of Arabia is still largely known from external rather than from indigenous sources. Some of the reasons for this are set out below.

---

<sup>4</sup> These dates are necessarily very approximate since the dating evidence for these inscriptions, almost entirely graffiti, is extremely slight.



**Figure 8.3:** Safaitic inscriptions at Jabal Says, southern Syria (C 25–32, see OCIANA). (Photograph by M.C.A. Macdonald)



**Figure 8.4:** Safaitic inscriptions on a stone at al-ʿĪsawī, southern Syria (C 3260–3264 see OCIANA). (Photograph by M.C.A. Macdonald)

### 8.1.1 Building a Digital Corpus: Challenges, Objectives and Perspectives

By the twenty-first century, the challenges faced by anyone trying to work with this material were considerable. Firstly, approximately 20,000 inscriptions had been published in scattered books and journal articles in many different languages. Additionally, an unknown number had been edited in unpublished dissertations, mainly in the Arab world, and many thousands were known to have been recorded but remained unedited and so unpublished.

Secondly, although Safaitic graffiti were first discovered in 1858, it was not until several decades later that some monumental ANA inscriptions were photographed or recorded by squeezes, and it was almost a century before epigraphic expeditions regularly photographed graffiti. Given that in the first 50 years the majority of texts were copied before the scripts had been deciphered, we are fortunate that, in general, the copyists were skilled, though they often made mistakes. Before the advent of digital photography, the number of films an expedition could take with it, and keep cool before and after use, was limited, so only a minority of inscriptions, particularly graffiti, was photographed. When editions were published, it was only possible to include a tiny number of photographs for reasons of cost. All this greatly hampered the progress of research into the languages, scripts, history and cultures of ancient North Arabia.

For this reason, there were virtually no research tools. The most recent grammatical sketch of the largest group (Safaitic) was published in 1943 (Littmann, 1943, pp. viii–xxiv), and of the second largest (Dadanitic) in 1954 (Caskel, 1954, pp. 60–77, repeated with minimal changes and corrections in Farès-Drappeau, 2005, pp. 61–77); there were no dictionaries, and the only list of names was published in 1971 (Harding, 1971).

Despite this situation, more and more inscriptions were recorded by Saudi and Jordanian academics and published in small handfuls, often with no photographs and little, often confused, information on their provenance.

If this situation was to be improved it was clear that all known inscriptions from ancient North Arabia needed to be sought out and brought together in a single, digital corpus, edited in a single international language to enhance access, with all available metadata, fully searchable texts and as many images as possible (preferably photographs). Between 1995 and 2003 Macdonald had begun a process of finding the sites in southern Syria where early travellers had copied Safaitic inscriptions, photographing the inscriptions, describing the sites and recording their location as accurately as possible (GPS was in its infancy and was forbidden in Syria at the time).

At the same time, with the help of Laïla Nehmé (CNRS Paris) and the late Geraldine King (independent scholar) he created a database of the Safaitic inscriptions (“The Safaitic Database”) using the platform, *4th Dimension*. This was maintained and expanded until 2012, when it was decided to use it as the basis for a corpus of *all* the inscriptions of ancient North Arabia, not just Safaitic. This was OCIANA, which was

based at the Khalili Research Centre, University of Oxford, and funded for three and a half years (September 2013 to March 2017) by the UK's Arts and Humanities Research Council (AHRC).

OCIANA's first objective was to identify all the known inscriptions of ancient North Arabia, whether published, edited in an unpublished dissertation, or recorded but unpublished. Ali Al-Manaser, a member of the project, among much else traced a large number of dissertations and sought and received permission from the authors and universities to include new editions of the inscriptions in OCIANA. Numerous scholars generously made available their photographs of published and unpublished inscriptions and gave the project permission to edit or re-edit them in OCIANA.

The second objective was to produce an up-to-date edition or re-edition of all the known inscriptions from ancient North Arabia in a single international language, English. This would be achieved by checking, and if necessary revising, a previous edition of the inscription in light of the most up-to-date knowledge. Previous readings and interpretations would be given in the *apparatus criticus* (translated into English if necessary), and there would be commentaries on the reading, translation or content of the inscription where necessary.

Thirdly, every word, name, genealogy, narrative, and prayer would be tagged to make it possible to search for all examples of words, grammatical features, expressions, personal, divine, place and group names, genealogies, etc.; something which by this time was becoming increasingly impossible to do in all the scattered publications on paper and online. This also meant that it would be easy to search across corpora to find, for instance, whether a particular word, grammatical feature or name is found in both Safaitic and Dadanitic, or Hismaic or Taymanitic. This, of course, provides the basis for the research tools that will be one of the major outcomes of the project (see below).

Fourthly, the bibliography of the inscriptions of ancient North Arabia was already large, with the added problem that scholars in the West had great difficulty hearing about the publication of books and articles produced in the Middle East, and to a lesser extent *vice versa*. Clearly, there was an urgent need for a regularly updated bibliography.

At the end of the second phase in March 2017, the database contained 42,672 (previously published and unpublished) inscriptions, the metadata of which had been entered and their data tagged over the course of three and a half years. Furthermore, over 100,000 negatives, prints and colour slides had been scanned and entered into the database.

The scripts bundled together under the heading "Thamudic" are only partially deciphered and it is therefore clearly necessary to find a way to enter the texts in a form which does not prejudice efforts to make a more satisfactory decipherment possible. It will therefore be necessary to develop a system of glyphs that, in a formalized fashion, imitate the shapes of the original glyphs in the inscriptions. This will then allow the exploration of repeated patterns of glyphs that is essential to the

decipherment of scripts. This is one of the prime aims of the third phase of the project. Once the decipherments have been made and thoroughly tested, the glyphs will be converted to roman script so that the “Thamudic” inscriptions can be searched along with all the other corpora.<sup>5</sup>

Having established the database and provided, for the first time, a fully-searchable corpus of most of the known inscriptions from ancient North Arabia, it is now time to use it to provide the basis for research into the languages, scripts, history and cultures of ancient North Arabia and to produce up-to-date research tools for the subject.

In the next phase, OCIANA will be used to produce concordances of words in context as the basis for the creation of online dictionaries and grammars. The *Dictionary of the Inscriptions of Ancient North Arabia* (DIANA) project based on OCIANA has already started this work (see section 8.3).

An up-to-date and easily updatable onomasticon will be produced both within each corpus and across all the corpora of the inscriptions. This will provide the material for a thorough study of the names within and across the various corpora.

Concordances of genealogies will also be produced. The Safaitic inscriptions, carved by nomads, form by far the largest corpus in the database (79%) and almost all provide genealogies varying from two to nineteen generations. The concordance will then be used not only to show the relationship of one author to another but, when combined with the provenance data of individual inscriptions, will be used to provide a picture of the movements of these nomads.

One of the most urgent and difficult problems to be faced is keeping the database up-to-date, given that several thousand previously unknown inscriptions are discovered each year, and that future research will inevitably require the revision of interpretations of individual texts. We are therefore working on the establishment of OCIANA within a university or other academic environment. It is hoped this will involve an endowment, making it possible to attract students and post-docs to continue work on the content, including both the editing of newly discovered inscriptions and the output of regularly updated research tools, as well as studies of the inscriptions and their contexts, and a constant updating of the bibliography.

## 8.2 The Development of OCIANA

In 2012, Michael Macdonald’s Safaitic Database was converted from the program *4th Dimension* to Filemaker Pro in preparation for Phase 2 of the OCIANA project. From September 2013, the OCIANA database was built in this application, although it also makes use of HTML, XML, SQL, and JavaScript. When planning for the development of the OCIANA database, we were aware that many projects in the field of digital

---

<sup>5</sup> For the technical problems that we faced and resolved see section 8.2.



epigraphy had opted to base the development of their databases on XML, specifically the TEI (Text Encoding Initiative)<sup>6</sup> and EpiDoc standards<sup>7</sup> (developed for tagging collections of text-based data). We chose not to base our new system on this format in the input process, but to ensure that any data within the database we were developing could be outputted as marked-up XML, for the purposes of Open Access, and opted to develop OCIANA in Filemaker Pro.

Whilst XML is a useful format for sharing data and data outputs, it suffers from being fairly unforgiving in terms of data entry, which will often mean that it takes considerably longer to enter and “mark up” content in platforms that use this standard. XML is also generally inefficient when dealing with very large sets of data, as it needs to load in the entire content of its file before users can start working on individual elements. Developers can create ways to work around these limitations, but the fact remains that XML is not really a database application, but rather a well-structured and delineated flat file of text. Its strength is its platform-independence, and the self-contained nature of its content, making it a great resource for data outputs; however, its strength does not extend to data management and manipulation, which is better managed by other applications and formats.

The benefits of using Filemaker Pro were considerable. It had proved to be a robust and stable platform, storing a great amount of data, serving a considerable number of concurrent users, with very low maintenance overheads.<sup>8</sup> Filemaker Pro could manage all the needs of the project, whilst ensuring that the corpus could be exported in formats that would be platform-independent and allow for sharing in line with Open Access standards.

Filemaker Pro offers an intuitive development environment, which allows for the rapid development of database solutions, whilst also offering a depth and flexibility that make it an extremely good fit for humanities data. A major benefit of the platform is that it is possible to publish Filemaker databases online with very little customisation, via the platform’s built-in web publishing engine.<sup>9</sup> The database had to be accessible and searchable via the internet, and the project’s AHRC funding stipulated that the data had to be freely available as an open-access resource, and its

---

<sup>6</sup> Further information about TEI is available online at [<http://www.tei-c.org>].

<sup>7</sup> Details about EpiDoc can be found at [<https://sourceforge.net/p/epidoc/wiki/Home/>].

<sup>8</sup> I have been working with Filemaker Pro since the late 1990s, and have used it to develop many different databases and applications: a Patient Information and Chemotherapy Management System for Cancer Research UK, used by the Medical Oncology Unit at Oxford’s Churchill Hospital for a period of almost ten years; and a number of databases for research projects at the Khalili Research Centre, University of Oxford.

<sup>9</sup> My work on the databases for the Pitt Rivers Museum, Oxford, had made use of this functionality, and the online catalogues have proved to be stable, and surprisingly speedy, even though the two catalogues, one for museum objects, and the other for photographic collections, each contained more than 250,000 records.

content reusable through Creative Commons licensing. Filemaker Pro allowed us to achieve this without the need for developing separate platforms from those used by the internal research team.

Central to the initial development of the database was the need to convert the flat-file structure of Michael Macdonald's Safaitic Database into a relational database, in order to handle efficiently the planned addition of an estimated 20,000 additional inscription records, as well as upwards of 100,000 images. In addition to these central requirements, the project team did not set out to create a simple searchable repository, but rather to build further functionality into the system, for the generation of research outputs, including the tagging of individual elements of inscriptions, such as nouns, verbs, adjectives, place names, personal names, divine names, and so on. The aim of this work was to lay the groundwork for grammars and dictionaries of these ancient languages, as well as searchable concordances of genealogies, and other important outputs that would help to shed light on the milieu from which the inscriptions came.

As the project progressed, it was possible to identify further tools and outputs we could develop on top of the core foundations we were building, and work to build them into the developing platform. I will touch on many of these over the course of this section, but the most significant development was the ability to output the entire corpus in pdf format, and its subsequent publication on the Bodleian Library's Online Research Archive (ORA). This dataset and publication is, at the time of writing, the single largest repository of data that is hosted on the ORA.

As outlined previously, the central core element, or record type, stored in OCIANA is the information about each individual inscription. This forms the basis for the whole database. It is possible to perform very detailed searches of the inscriptions contained in OCIANA, and each inscription record contains a large amount of information, split into a series of fields.

Clearly, the main element of any inscription record is the text of that inscription, and this textual content is recorded in both transliteration and translation within OCIANA. Transliterations are presented both in roman characters and in glyphs imitating the original letter forms,<sup>10</sup> with the customary editorial apparatus. The translations, *apparatus criticus*, commentaries and all information about each inscription are presented in a single international language, English, regardless of the language in which the inscriptions were first edited.

Moreover, alongside the text, it was necessary to record many related items of metadata. Whilst every inscription in the database has a unique OCIANA identifier, it will also have a siglum assigned to it. This will usually indicate the original publication or survey from which the inscription hails, as well as providing some

---

<sup>10</sup> This was done for those used to non-roman scripts, such as Arabic, many of whom have said that they find it easier to read these glyphs than the transliteration into roman letters with diacritical marks.

indication of its geographical provenance. These are elements that would be difficult to replicate via a system of database-generated sequential numbering. Inscriptions that had already been published were known by established reference sigla, such as those published in Winnett & Harding, 1978 (WH); Littmann, 1943 (LP); and from the *Corpus Inscriptionum Semiticarum Pars V* (C).<sup>11</sup> Where we were entering previously unpublished inscriptions we opted to follow this model, and created new sigla to assign to these collections.

Inscription records also contain several other fields containing research notes specific to that inscription, such as an alternative siglum, commentary and *apparatus criticus*. However, a large number of other fields for each inscription record contain data common to large groups of inscriptions, and many of these fields are therefore connected via a relational model to other database tables within OCIANA.

**Table 8.1:** Some of the fields assigned to each inscription in the database, with links to other tables within OCIANA

Field	Role	Linked?
inscription_recordID	The auto-generated unique identifier for each record	Yes
inscription_siglum	The historical reference identifier for each record	No
inscription_script	The script that the inscription is written in	Yes
inscription_fullText	The transliterated text of the inscription	No
inscription_translation	The English translation of the inscription	No
inscription_appCrit	The <i>apparatus criticus</i> of the inscription record	No
inscription_commentary	Research notes and commentary on the inscription	No

In the database's table of inscriptions there are, of course, many more fields than those listed in Table 8.1 (146, in fact), with some fields containing legacy data from the Safaitic Database, and others applying calculations and alterations to other fields. For example, the "inscription\_fullText" field allows the text of an inscription to be input complete with editorial marks indicating where some of the glyphs are illegible, scratched out, or otherwise uncertain, but a second field removes this mark-up information in order to allow users to search the inscriptions without these being included. In addition to these further fields, a number of fields focus on the internal administration of records, including the date a record was created, when it was last modified, which researcher entered the data, and so on. A final set of calculation fields allows the database to output its content in HTML, for the publication of unique web pages (every inscription in the database has its own static web page, allowing for indexing by Google and other search engines), as well as XML. This facilitates sharing of the data with researchers via the Bodleian Online Research Archive.

<sup>11</sup> See Ryckmans, 1950–1951.

As mentioned earlier, the table of inscriptions is one of several tables in the database, and is linked via several internal relationships to a series of other tables. These tables serve a number of important functions, including links to images of inscriptions, bibliographic references, and other data common to the corpus. Relational table links are particularly important for one of the key functions of OCIANA as a research tool, which is the functionality that allows editors to tag elements of each inscription. Individual elements of inscriptions within the corpus contain a number of differing functions or characteristics. Individual words may be tagged as grammatical elements, and names are tagged for genealogical and onomastic searches. Collections of words or phrases may be tagged as narrative elements or prayers, and genealogies are also tagged. When editors select a word or phrase, they then apply an appropriate tag to their selection, and this will then create a new related record in the appropriate supporting table within the database, allowing us to create tables of unique grammatical content, genealogies and genealogical concordances, and a detailed list of onomastic elements within inscriptions. These supporting tables have then allowed us to create comprehensive word lists, complex concordances of genealogies and words, and to work towards creating grammars and dictionaries for the scripts and languages contained within OCIANA.

Not all of this functionality is yet available to online users of OCIANA, but they still have access to the entire corpus of inscriptions, and the ability to search its content, and the contents of a number of supporting tables, in great depth.<sup>12</sup>

Online users of OCIANA can freely search all of the published inscriptions (42,672 in total), the table of tagged grammatical elements (123,062), the tagged onomastics (95,673), and both the tagged genealogies and their concordances (37,222). As mentioned earlier, each individual inscription has its own unique web page containing all the information about it, details of its tagged elements, a list of the bibliographic references, and all available photographs of that inscription. The citation URL is listed on the page for each record, and we would encourage anyone making use of OCIANA to include this in their publications. An example of an inscription record from the online database is shown below, with the URL for citation and linking to the record indicated at the end. The example does not show the glyphs or images here, and the list of tagged grammar and onomastics has been omitted.<sup>13</sup>

---

<sup>12</sup> The online version of the OCIANA database can be freely accessed at [<http://163.1.184.24/fmi/webd/OCIANA>], and an overview of the online functionality of the database is covered in a talk I gave at the Digital Humanities Summer School (DH@OxSS) in July 2015, which can be viewed at [[http://krc.orient.ox.ac.uk/resources/ociana/ociana\\_dhoxss.mp4](http://krc.orient.ox.ac.uk/resources/ociana/ociana_dhoxss.mp4)].

<sup>13</sup> The complete record, including these elements can be viewed at [[http://krc.orient.ox.ac.uk/ociana/corpus/pages/OCIANA\\_0033109.html](http://krc.orient.ox.ac.uk/ociana/corpus/pages/OCIANA_0033109.html)].

**Sigla:** AH 001; Sima 1999: 35–36; D 134 **Script:** Dadanitic **Language:** Dadanitic

### Transliteration

- 1: *bn[w]d/w whb'm/w* '–
- 2: *wd/w lb'n/bnw*
- 3: *s'd'l/d yf'n/z*–
- 4: *llw/zll/h- nq/l*
- 5: *dġbt/frd -hm*

### Translation

- 1: {Bnwd} and Whb'm and '–
- 2: wd and Lb'n sons of
- 3: S'd'l of the lineage of Yf'n per–
- 4: formed the zll-ceremony of the top of the mountain for
- 5: Dġbt and so favour them

### Apparatus Criticus

#### TEXT

Line 1: Abū l-Ḥasan followed by Farès-Drappeau: *bnd w* rather than *bn[w]d*; Sima: *w-'tb'm* rather than *w- whb'm*. The latter is clear on the photograph, although the *h* is slightly damaged and was copied and read as *t* by Abū l-Ḥasan.

Line 2: Abū l-Ḥasan: *wm* for *wd*; Sima: *wg* for *wd*.

Lines 3–4: Abū l-Ḥasan: 'ʔllw ʔll for 'zllw zll.

#### TRANSLATION

Lines 3–4: 'zllw h- zll, Sima: 'they covered ???'; Farès-Drappeau: '(they) offered the sacrifice'.

Line 4: *h- nq*, Abū l-Ḥasan: 'the female camels'; Sima does not translate it; Farès-Drappeau: 'the female camel'.

#### DISCUSSION

Hidalgo-Chacón Díez 2016: 128, for the divine name *Dġbt*.

### Commentary

The restored [*w*] in the first personal name is based on the existence of the personal name *Bnwd* in the inscription AH 011/1. A root *bnd* has not been found in Semitic (Cohen et al. 1970–: 71). The translation of the construct phrase *zll h-nq* is based here and in other texts on interpreting *h-nq* either as a place name or as a common noun from the Arabic word *nīq* 'mountain-top'. Given that the inscriptions mentioning *h-nq* are located on the way up Ġabal 'Ikmaḥ or at the top of Ġabal Umm Daraġ, it seems unlikely that they would be recording the sacrifice of female camels (*nāq* or *nūq*), as suggested by Abū l-Ḥasan and Farès-Drappeau (see the *apparatus criticus*).

**Subjects:** Genealogy Lineage Religion Deity Prayer Topographic features

**Country:** Saudi Arabia

**Region:** Al-Madīnah

**Site:** Oasis of al-‘Ulā

**Latitude:** 26.616667

**Longitude:** 37.916667

**Present Location:** In situ

**Notes:** Al-‘Uḏayb (Ġabal ‘Ikmaḥ)

**References:**

[AH] Abū ‘l-Ḥasan, Ḥ. ‘A.D. *Qirā’ah li-kitābāt liḥyāniyyah min ġabal ‘akmaḥ bi-miṭṭaqaṭ al-‘ulā*. Al-Riyāḏ: Maktabat al-malik faḥd al-waṭaniyyah, 1997. Pages: 53–61 Plates: 1

Cohen, D., Bron, F., Lonnet, A. *Dictionnaire des racines sémitiques: ou attestées dans les langues sémitiques: comprenant un fichier comparatif de Jean Cantineau*. Paris: Mouton (fasc. 1-2)/Leuven: Peeters (fasc. 3–), 1970–. Pages: 71

[D] Farès-Drappeau, S. *Dédan et Liḥyān. Histoire des Arabes aux confins des pouvoirs perse et hellénistique (IVe–IIe s. avant l’ère chrétienne)* (Travaux de la maison de l’Orient 42) de la maison de l’Orient, 42). Lyon: Maison de l’Orient et de la Méditerranée — Jean Pouilloux, 2005. Pages: 212

Hidalgo-Chacón Diez. M. del C. The divine names at Dadan: a philological approach. *Proceedings of the Seminar for Arabian Studies* 46, 2016: 125–136 Pages: 128

Sima, A. *Die lihyanischen Inschriften von al-‘Uḏayb (Saudi-Arabien)*. (Epigraphische Forschungen auf der Arabischen Halbinsel, 1). Rahden/Westf.: Leidorf, 1999. Pages: 35–36

**URL of this record (for citation):**

[http://krc.orient.ox.ac.uk/ociana/corpus/pages/OCIANA\\_0033109.html](http://krc.orient.ox.ac.uk/ociana/corpus/pages/OCIANA_0033109.html)

In Phase 3 of the project, we intend to enhance the database in a number of ways, with perhaps the most interesting development relating to the group of inscriptions known as “Thamudic”. We will need to work on them by allowing the entry of inscriptions in glyph format, and then provide researchers with the ability to assign transliterations in roman characters to these glyphs at a later date, as they work to complete the decipherment of these scripts. Additional work would include moving some of the provenance data into separate related tables, as the work completed in Phase 2 has allowed us to develop a list of important sites of inscriptions in the region. At present, we are still at the early stages of planning for Phase 3, but we hope to begin work on the next stage of developments in 2018.

### 8.3 The Future of OCIANA

Phase 3 of OCIANA has three goals. The first is to keep the database up to date. The rapid pace of discovery requires the constant entry of new inscriptions and bibliography to ensure that it can be used to the maximum degree as a research tool. Hundreds of new Ancient North Arabian inscriptions are published each year, often in difficult-to-access publications of Middle Eastern universities or unpublished Master's and Ph.D. theses from Jordan and Saudi Arabia. Not only must these new texts be sought out, but also the photographs must be scanned, readings verified, along with the standard insertion of metadata.

In Phase 3, OCIANA will also fill an important gap in the database's current documentation: "Thamudic". This is a pending category covering the various corpora of Ancient North Arabian inscriptions that have not yet been subjected to thorough study. Thamudic now includes four categories: B, C, D, and F,<sup>14</sup> spanning from Syria in the north to Yemen in the south. What is more, each of these classifications includes a remarkable amount of variation in letter shapes and in some cases the identification of a glyph with a phoneme is unclear – in other words, some of the scripts in the "Thamudic pending file" have not yet been fully deciphered. This fact presents a challenge to inserting data into OCIANA, in particular when it comes to transliteration. Rather than transcribing the glyphs by their assumed phonemic equivalent in the roman alphabet, of which we are often unsure, the transliteration of the Thamudic material will encode the actual letter shapes themselves. To illustrate, the glyph # represents *ḏ*, *ḏ*, and *ṭ* in Ancient South Arabian, Safaitic, and Hismaic, respectively. Rather than assuming one of these values in a poorly understood Thamudic text, we will simply encode the glyph with a standardized version of the glyph itself. This neutral representation will then allow the researcher to revisit the patterns of distribution of problematic letter shapes across the entire corpus, allowing for a more precise classification of scripts and ultimately a clearer understanding of these enigmatic corpora. It is anticipated this process will eliminate, or greatly reduce, the "Thamudic" pending category and permit the recognition of new, properly understood, scripts.

The third goal of Phase 3 is the use of OCIANA as a research tool. The first sub-project of this goal is currently being realised as the *Dictionary of the Inscriptions of Ancient North Arabia* (DIANA), an online open-access supplement to the OCIANA database. The dictionary will include every lexical item contained in OCIANA. Each word will have a dedicated entry, with a full etymological discussion, ample

---

<sup>14</sup> Some of the "Thamudic F" or "Southern Thamudic" graffiti have recently been deciphered by C.J. Robin and have been removed from the "Thamudic pending file" and relabeled "Himaitic" from the area in southern Saudi Arabia where they are found. See Robin & Gorea, 2016.

illustrative examples, as well as synonyms. Users can easily follow a link to OCIANA to see all the attestations of a given lexical item in the corpus.

Sample Safaitic entry:

**ts<sup>2</sup>wq** *v.t2-stem. to long for; to feel longing.* *Root: s<sup>2</sup>wq.* [tas<sup>2</sup>awwaqa] [tas<sup>2</sup>weqa] HCH 44: ts<sup>2</sup>wq 'l-b-h w 'l-'ḥt-h 'he longed for his father and for his sister'; C 95 wgd s<sup>1</sup>fr dd-h f ts<sup>2</sup>wq 'he found the inscription of his paternal uncle and so was filled with longing' *Variant: ts<sup>2</sup>wqw.* [tas<sup>2</sup>awwaqaw(?)] RSIS 204: ts<sup>2</sup>wqw 'l-ṣḥ 'he longed for Ṣḥ' *Variant: ts<sup>2</sup>yq.* [tas<sup>2</sup>ayyaqa] KRS 124: ts<sup>2</sup>yq l-ḥbb 'he longed for a friend' *Third feminine singular:: ts<sup>2</sup>wqt.* [tas<sup>2</sup>awwaqat] Damascus Museum 2786 = RyDamas 5537: l PN w ts<sup>2</sup>wqt 'l-nmn 'by PN and she longed for Nmn' *Note: His: ts<sup>2</sup>wq (CH.R716) || The equivalent of CAr ištāqa 'ilay-hi 'he was, or became, desirous of it ... [or he longed for it in his soul]' (Lane, 1620b).<sup>15</sup>*

Other sub-projects of this Phase, pending funding, include the creation of an up-to-date onomasticon, and in-depth studies of the individual Ancient North Arabian corpora, such as Hismaic, the various properly identified scripts emerging from the study of “Thamudic”, and the minor corpora.

Currently, DIANA and the maintenance of the database are progressing, but in order for phase 3 to be fully realized, the team is preparing applications for funding and institutional support, both at Oxford and Leiden University.

## Bibliography

- Farès-Drappeau, S. (2005). *Dédan et Lihyān. Histoire des Arabes aux confins des pouvoirs perse et hellénistique (Ive–Ile s. avant l'ère chrétienne)* (Travaux de la maison de l'Orient, 42). Lyon: Maison de l'Orient et de la Méditerranée - Jean Pouilloux.
- Caskel, W. (1954). *Lihyan und Lihyanisch* (Arbeitsgemeinschaft für Forschung des Landes Nordrhein-Westfalen, Geisteswissenschaften 4). Köln: Westdeutscher Verlag.
- Harding, G.L. (1971). *An Index and Concordance of Pre-Islamic Arabian Names and Inscriptions* (Near and Middle East Series 8). Toronto: University of Toronto Press.
- Lane, E.W. (1863–1893). *An Arabic-English Lexicon, Derived from the Best and Most Copious Eastern Sources*. London: Williams & Norgate.
- Littmann, E. (1943). *Safaitic Inscriptions. Syria. Publications of the Princeton University Archaeological Expeditions to Syria in 1904–1905 and 1909. Division IV. Section C*. Leiden: Brill.
- Macdonald, M.C.A. (2008). The Phoenix of Phoinikēia: Alphabetic reincarnation in Arabia. In J. Baines, J. Bennet, & S. Houston (Eds.), *The Disappearance of Writing Systems: Perspectives on Literacy and Communication* (pp. 207–229). London: Equinox.
- Macdonald, M.C.A. (2010) Ancient Arabia and the written word. In M.C.A. Macdonald (Ed.), *The development of Arabic as a written language* (Supplement to the Proceedings of the Seminar for Arabian Studies volume 40) (pp. 5–27). Oxford: Archaeopress.
- Overlaet, B., Macdonald, M.C.A., & Stein, P. (2016). An Aramaic–Hasaitic bilingual inscription from a monumental tomb at Mleiha, Sharjah, UAE. *Arabian Archaeology and Epigraphy*, 27, 127–142.

<sup>15</sup> The sigla of the inscriptions can be found in OCIANA.



- Robin C.J. & Gorea M. (2016). L'alphabet de Ḥimā (Arabie séoudite). In I. Finkelstein, C.J. Robin, & T. Römer (Eds.), *Alphabets, texts and artifacts in the ancient Near East. Studies presented to Benjamin Sass* (pp. 310–375). Paris: Van Dieren.
- Ryckmans, G. (Ed.). (1950–1951). *Corpus Inscriptionum Semiticarum. Pars V. Inscriptiones Saracenicis continens, Tomus 1. Inscriptiones Safaiticae*. (2 volumes). Paris: Imprimerie nationale.
- Stein, P. (2005a). The Ancient South Arabian minuscule inscriptions on wood: A new genre of pre-Islamic epigraphy. *Jaarbericht van het Vooraziatisch-Egyptisch Genootschap 'Ex Oriente Lux'*, 39, 181–199.
- Stein, P. (2005b). Stein vs. Holz, *musnad* vs. *zabūr* — Schrift und Schriftlichkeit im vorislamischen Arabien. *Die Welt des Orients*, 35, 118–157.
- Stein, P. (2013). Paleography of the Ancient South Arabian script. New evidence for an absolute chronology. *Arabian Archaeology and Epigraphy*, 24, 186–195.
- Winnett, F.V. & Harding, G.L. (1978). *Inscriptions from Fifty Safaitic Cairns* (Near and Middle East Series 9). Toronto: University of Toronto Press.

Anne Multhoff

## 9 A Methodological Framework for the Epigraphic South Arabian Lexicography. The Case of the Sabaic Online Dictionary

**Abstract:** This paper describes the concept and functionalities of an online reference dictionary for Sabaic, aiming to present all extant lexical material of this Ancient South Arabian language. After introducing the features of the corpus, several methodological issues, and the solutions adopted within the project are illustrated, focusing in particular on the annotation of morphological analysis (treatment of ambiguous forms, homographs, heterographs with identical meaning, variant readings, incorrect forms). The conception developed to present the material online is also described.

**Keywords:** Sabaic, Ancient South Arabian, lexicography, variants, translations

### 9.1 Introduction

#### 9.1.1 General Remarks

Over recent decades, a huge amount of new Ancient South Arabian inscriptions has come to light. This has been published in collections such as Arbach & Schiettecatte, 2006 or Priolella, 2013 to name but a few, but mostly in scattered editions comprising only a few texts. This material not only contains hitherto unknown lexemes, but also calls for a reconsideration of quite a number of well-known terms in the Sabaic lexicon. The available dictionaries on Sabaic such as Beeston et al., 1982 or Biella, 1982 thus no longer reflect the present state of research. The same holds true for dictionaries on other Ancient South Arabian idioms such as Qatabanic (Ricks, 1989) and Minaic (Arbach, 1993). Moreover, apart from a considerable quantitative increase of the material, we are also confronted with a qualitative leap, as completely new text genres have emerged, particularly among the everyday correspondence on wooden sticks (cf. Stein, 2010 & Maraqtan, 2014). Though revision of at least part of the material included in the extant dictionaries was rightly demanded in both well-meaning (e.g. Lundin, 1987) and cynical reviews (cf. Jamme, 1985, pp. 202–269), a revised second edition of Beeston et al., 1982 never appeared. An up-to-date presentation of the

---

Anne Multhoff, Friedrich-Schiller-Universität Jena



© 2018 Anne Multhoff

This work is licensed under the Creative Commons Attribution 4.0 International license (CC BY 4.0)

lexical data gathered over the past 30 years is therefore clearly a desideratum of the scientific community, both within and outside Ancient South Arabian Studies proper.

### 9.1.2 Scope of the Project

The project described in the present paper<sup>1</sup> aims at a reconsideration of the lexical material. It will result in a reference dictionary (“Belegwörterbuch”) that will include a complete lexical survey of the Sabaic material published so far. In contrast to other projects on Ancient Arabian epigraphy featured in this volume, such as DASI and OCIANA, it is not focused on the epigraphic corpus as such, but uses the latter as a basis for lexicographic work. Digitization of material is thus not considered as a result intended for public use, but rather as a practical means to collect and organize large amounts of data.

The application is running on a Microsoft-Windows-Server on which the IIS (Microsoft) is installed as internet server. The data is stored on an instance of the Microsoft SQL Server Express. Applied programming languages are C# and JavaScript. While the internal working platform was designed in ASP.Net, the more modern ASP.Net MVC is used for the public web presence. Furthermore, the JavaScript framework JQuery is used in the web presence.

Two different concepts, adjusted to the various parts of the working process, are used for data management. First, a collection of the epigraphic material is needed as a material base to reference lexemes. For the annotation of texts an XML format was chosen. An editing view of each annotated text is generated from the XML document as a HTML view via a JavaScript routine. Annotations are directly assigned to words as XML attributes. XML documents thus generated are stored in the database. Following the grammatical analysis, the information contained within these XML documents is divided into the tables of the relational database to enable further processing.

Second, a complete set of the interpretations given in the literature is collected. This is to facilitate lexicographic work, comprising translations from text editions, extant dictionaries and further lexicographical material published both in compendia (e.g., Sima, 2000) and specialized articles (e.g., Robin, 2013), or studies on Ancient South Arabian culture and history (e.g., Beeston, 1976) to name but the most important genres. The collected material is further enriched with etymological parallels from other Semitic languages. This material has no intrinsic relation to a specific inscription. It is directly stored in the tables of the relational database.

---

1 [<http://sabaweb.uni-jena.de>].

## 9.2 Material Base

### 9.2.1 Character of Material

Sabaic is part of a group of several interrelated epigraphically documented Semitic languages, commonly referred to as Ancient South Arabian, which were spoken in the territory of modern Yemen from the early 1<sup>st</sup> millennium BCE up to the 6<sup>th</sup> century CE (cf. Stein, 2011). The material, however, is rather extensive both in respect to absolute length of attested texts and as far as lexical variance is concerned, at least if compared to contemporaneous European texts of equal genre, as e.g. Latin inscriptions. Reading is considerably facilitated by the regular use of word dividers<sup>2</sup>. As with most Semitic languages, the script is highly defective: basically only consonants are noted, with the only exception of final long vowels.<sup>3</sup> Abbreviations are absent. Inscriptions as material objects are written on durable material, mostly stone or metal. They thus constitute a primary source that may be damaged or destroyed, but is not prone to textual alterations. However, this only applies to the object as such. A comparatively large portion of the Ancient South Arabian material is only known from copies or transcriptions of various qualities, made by modern scholars. As photographic documentation, though already requested in the review literature at a comparatively early time (cf. Schlobies, 1936, p. 58, n.1), is often lacking or of poor quality, the actual appearance of these inscriptions can no longer be checked. This material contains a certain amount of corrupt forms, including obvious faults (cf. Stein, 2002, esp. the exhaustive examples pp. 447–452). The latter are often corrected by later reeditions based on photographs or rediscovered originals.<sup>4</sup> For a large amount of poorly published material there is still no reliable documentation available. Nevertheless, rectification of obvious faults was often undertaken by later editions. As a result, these inscriptions present a certain amount of variant readings that can rather be compared to manuscripts.

---

<sup>2</sup> These function in a roughly similar way to spacing in modern Arabic, i.e. there is a certain range of prefixes, mainly monoliteral particles, and suffixes, mainly pronominal suffixes, that are supposed to form a continuous string of characters with the main word.

<sup>3</sup> This only applies to the very end of a string of characters (see above), long vowels before suffixes always being suppressed in script.

<sup>4</sup> A particularly prominent example is Ir 13 published by al-'Iryānī (1973, pp. 74–86) on the basis of a highly defective copy. A portion of the text has been rediscovered and is reedited by Arbach (2001). Certain grammatical and lexical oddities disappeared in the course of this reedition, e.g. an ungrammatical infinitive *qtq* in § 5 that turned out to be a simple *qtl* (cf. the reedition, line 4). Note that the misread form is included in both Beeston et al. (1982, p. 110) and Biella (1982, p. 455), and has thus become part of lexicographical discussion.

### 9.2.2 Collection of Material

Ancient South Arabian inscriptions are scattered over a wide range of different publications which may comprise huge collections of different material (e.g., CIH and RES) and full inventories of excavations (e.g., Jamme, 1962), but may also focus on individual texts or passages. An exhaustive collection of material is undertaken by DASI, but is not yet completed for Sabaic. The compilation of material underlying the present dictionary already started back in the late 1990s, then still in a DOS format. This collection consisted mainly of analytic transcriptions and bibliographic information. The material was originally meant as a basis for grammatical studies and was thoroughly prepared for this purpose. Information on Semitic roots or fuller, non-assimilated forms was thus encoded. Since the latter are also important for lexicographic work, the compiled file, subsequently completed and augmented by newly published material, was considered an ideal base for a dictionary. The data is by now converted in an XML format.

### 9.2.3 Organisation of Material

The dictionary has a modular approach. The Sabaic material is split into several sub-corpora (such as votive inscriptions, building inscriptions, juridical texts etc.), which are processed separately. For the time being, the dictionary includes major parts of the votive inscriptions, which actually form the most comprehensive genre among the Sabaic texts (an up-to-date account of the incorporated material can be found on the website).

Within these corpora, lexicographic work is organized according to inscriptions, i.e. all lexical items of a given text are considered, irrespective of alphabetical order. Work thus started with a micro-dictionary containing the vocabulary of a single inscription, this core being constantly enlarged with material from other texts. Therefore, not only does the actual number of lexemes increase over time, but also their extent.

The project aims to present all extant lexical material. This also includes probable or even actual faults. Variant readings and – to a lesser extent – unusual or even faulty orthographies are thus included in the basic text.<sup>5</sup> However, uncertain and incorrect readings are marked as such. A reconstruction of “correct” texts is intended for presentation purposes, but often turns out to be impossible. This is, in most cases,

---

<sup>5</sup> Note that the textual structure of passages with variants is fairly complicated in respect to encoding. This unfortunately leads to unexpected and unwanted results that also effect the presentation in the website.

due to deficient editions lacking proper documentation (cf. section 9.2.1), but may also result from the limited range of our knowledge, especially in damaged contexts.

### 9.3 Morphological Analysis

The morphological analysis of texts is performed by manipulating the XML document. In the process of tagging a word, information is stored as XML attributes to the corresponding XML tag. These attributes provide information on the actual lemma (e.g. *bn* “son” vs. *bn* “bān-tree”, preposition *bn* “from” etc.), its lexical category and grammatical subcategories (if appropriate). Lexical categories include noun, pronoun, verb, preposition, conjunction, other particle and other.<sup>6</sup> To facilitate a clear presentation of the lexical material, names<sup>7</sup> and fragments<sup>8</sup> are treated as separate categories. Since this classification is part of the lemma, as are roots and meanings, it is substituted automatically once the correct lemma is chosen. In ambiguous contexts, multiple tags can be assigned. However, this possibility is kept to a minimum to avoid confusion of the reader.

On the other hand, grammatical subcategories such as gender, number, state, conjugation and the like, are specific to the actual word. Since the defective Sabaic orthography includes many ambiguous forms<sup>9</sup>, these tags are edited manually, based on the particular context of the word. If neither form nor context allows a clear identification of categories, forms are tagged as “unspecific”.<sup>10</sup> Morphological tags are used to create a morphological catalogue of attested forms, which is a vital part of the reference dictionary.

---

<sup>6</sup> This category – in reality almost empty – was created to avoid technical problems with unexpected cases.

<sup>7</sup> These show substantial differences from “normal” lexemes. While forms like *yzd* or *yhn'm* expose a difference between (verbal) morphological form and (nominal) syntactical content, composite names like *krb-'l* or *whb-'wm* cannot be assigned to a single root. A full inclusion of names would thus destroy the otherwise exclusive relation between lemma and root. Furthermore, a thorough analysis of onomastics is not in the scope of the present project. Nevertheless, different categories of names (such as anthroponyms, theonyms, toponyms etc.) are differentiated so as to facilitate further studies on this material.

<sup>8</sup> This only applies to forms that cannot be reconstructed with a sufficient degree of certainty. A comprehensive study of the latter is not appropriate at the given state of research. Furthermore, minor fragments, often only consisting of a single initial or final letter preserved, are highly ambiguous and do not constitute lexemes in the proper sense of the word.

<sup>9</sup> Thus, a form like *hmr-hw* can be interpreted as an infinitive “to grant him” and as a finite form “he granted him” or “they granted him”.

<sup>10</sup> A general use of multiple tags in these often fragmentary contexts would lead to a rather unwelcome artificial increase of the catalogue of forms and thus does not allow adequate presentation.

Furthermore, tagging provides information on reliability of attestation: certain, uncertain, supplemented<sup>11</sup> or wrong<sup>12</sup>. While both uncertain and wrong forms are marked as such,<sup>13</sup> supplemented forms are simply excluded from presentation.

## 9.4 Definition of Lemmata

The definition of lemmata follows practical considerations. Forms are thus treated as separate items if their respective contexts show sufficient differences to consider them as such. Sabaic being an extinct language, this is more or less the approach of all extant lexicographic literature in the field.

### 9.4.1 Treatment of Homographs

Homographs belonging to different grammatical categories are commonly treated as separate items in the scientific literature (cf. the organization of material in both Beeston et al., 1982 and Biella, 1982). Obviously, a noun like *qrm* “garrison” should be differentiated from the homographic verb *qrm* “to be on garrison duty”. The difference between such morphological categories is, however, sometimes difficult to observe in existing contexts. Distinguishing between infinitives of the base stem, following a pattern *fʿl*, and abstract nouns showing a similar pattern in the construct state, is a particularly delicate case.<sup>14</sup> Particularly in stereotype contexts, the grammatical form of a certain lexeme may resist disambiguation.

The situation is even more complicated for homographs belonging to the same grammatical category, i.e. homograph verbs or homograph nouns. These were probably differentiated via vocal patterns, a means that is not featured in the defective script (for nominal forms cf. Stein, 2003, esp. pp. 56–62). Furthermore, the rather complicated system of verbal stems and their eventual graphic distinction was only fully understood in the last decade (cf. Multhoff, 2011). Consequently, Sabaic lexicography has not yet developed consistent standards to deal with this material. Delimitation of lexemes in existing dictionaries is thus often rather arbitrary,

---

**11** Forms are generally classified as “supplemented” if no single radical is preserved.

**12** This includes both misreadings by modern editors and actual mistakes by the Sabaean writer. A further specification, though desirable as such, is in many cases impeded by a lack of reliable (photographic) documentation.

**13** Uncertain readings are marked with a minor question mark (usually only if the reading is certain neither from script nor from context); incorrect forms and identifications are crossed out.

**14** An example is *šsy* “upheaval” in the expression *nqʿ w-šsy šnʿm* “elevation and upheaval of an enemy”: the word could well be defined as an infinitive of the equally attested verb *šsy* of more or less equivalent meaning.

sometimes even summarizing different graphemes under one single lemma.<sup>15</sup> Given the progress in our understanding of Sabaic verbal stems over the last fifteen years, it is now possible to distinguish different verbal stems. All stems show at least occasionally unequivocal forms, mostly in infinitives. Even though these are not always attested, the fundamental disambiguation of the system has yielded a set of semantic criteria that often enables definition in otherwise uncertain cases. Verbal homographs can thus normally be clarified, presenting different stems as different lexemes and identical stems as singular lexemes, as with *ʿrb* 0<sub>1</sub> “to enter” besides 0<sub>2</sub> “to offer” (but see below, section 9.4.2).<sup>16</sup>

The situation is more complicated with nominal forms. If we compare related languages such as Arabic, a much bigger amount of homographs is to be expected, but clear morphological or semantical criteria are missing. In this particular case, contexts are carefully checked. Forms appearing in similar contexts are generally treated as a single lexeme. Other forms are considered as such if there are no convincing arguments against this assumption. On the other hand, forms are split into different lexemes if clear semantic differences appear from constructions or contexts, as in the case of *dhb*, a homograph comprising lexemes with the meaning “bronze”, “oasis, irrigated area”, “irrigation, flood-water” and a certain “measure of capacity”.

#### 9.4.2 Deliberate Splitting of Lexemes

In certain cases, single lexemes are deliberately split up. The most common cases are particles. Conjunctions are generally separated from homograph prepositions. Very frequent particles such as *b-*, *l-* and the pronoun *q-* are further divided into different semantic or contextual categories to enable a clear presentation.<sup>17</sup> Verbal and nominal lexemes are split if they are clearly derived from different forms.<sup>18</sup> Lastly, a small

---

<sup>15</sup> Cf. e.g. the merge of *hḏrʿ*, *tḏrʿ* and *stḏrʿ* into a single entry that can be observed in Beeston et al., 1982, p. 42. A critical re-evaluation of this approach can already be found in the review by Lundin (1987, p. 49): “As a result, the *lemmata* [sic. l.] are presented in far too abstract a way without distinguishing words with similar meanings and grammatical forms (i.e. verbal stems).”

<sup>16</sup> However, uncertain cases such as morphologically ambiguous forms in divergent contexts often have to be considered as separate items.

<sup>17</sup> This is often done in dictionaries containing selected examples (as is the case with Biella, 1982), but avoided in more concise presentations such as Beeston et al., 1982. However, this approach proved rather dysfunctional in practice for most lexemes concerned in the context of a full citation of references.

<sup>18</sup> Thus *hwfy* is split into denominal “to give well-being” (from *wfy* “well-being”) and deverbal “fulfil, hand over” (from *wfy* “to belong to”).



number of nouns in ubiquitous contexts are given separate entries to allow a clearer presentation of the remaining forms in the context of a fully referenced dictionary.<sup>19</sup>

### 9.4.3 Heterographs with Identical Meaning

Sabaic shows a certain number of words that are different in form, but apparently similar in meaning. This mainly applies to a) different rendering of weak radicals (*w* or *y*) and b) otherwise identical nouns with and without final *-t*. While some of these forms can be explained by diachronic or regional variation,<sup>20</sup> the motivation of other variant forms is less clear.<sup>21</sup> Those may represent different lexemes, but may also refer to different numbers (the distribution of which rests equally unclear). Different graphemes are considered as one single lemma if their relation is clearly grammaticalized, as is the case with diachronic or regional variation. Other cases are mostly treated as different lexemes.<sup>22</sup>

### 9.4.4 Treatment of Incorrect Forms

The published Sabaic material, accumulated over a period of almost 150 years, comprises a surprisingly high number of incorrect forms. These include both actual faults of the Sabaean writer (e.g. a merger of similar characters, as in *'rz* instead of *'rḏ* “earth”) and misreadings (sometimes even misspellings) by modern editors. In the absence of reliable photographic documentation, it is difficult, if not impossible, to attest to the correctness of a given form.

However, misreadings have sometimes become real “classics” over a certain period of time and could have provoked a rather huge amount of material, both translations and further reflections, and even found their way into the extant dictionaries.<sup>23</sup> A simple exclusion of these ghost-words from the dictionary will probably not solve the problem of their constant reappearance in (especially non-specialist) literature. On

<sup>19</sup> This is the case with *b'l* “lord” in the phrase “NN [theonym], the lord of NN [toponym]” and *mlk* “king” in the phrase “NN, the king of Saba’ [and ...]”. In the latter phrase, all ethnonyms concerned, (i.e. Saba’, *ḏū-Raydān*, *Ḥaḏramawt* and *Yamanat*) are also split up.

<sup>20</sup> Cf. e.g. verbal roots III *w* that are generally rendered as III *y* in Southern Sabaic dialects and late Sabaic texts (the latter actually being an offshoot of South Sabaic), thus *ḏky* besides *ḏkw* “to send; to expel”.

<sup>21</sup> Cf. e.g. *b's* besides *b'st*, both meaning “evil, malice”.

<sup>22</sup> In some uncertain cases, however, multiple tags may be used to link several possible lexemes to one single form.

<sup>23</sup> Cf. e.g. *\*mhrk* “booty” featured in Beeston et al., 1982, p. 57, and Biella, 1982, p. 117, misread from damaged *mhrḡ*.

the other hand, assumed scribal faults are not always as obvious as the example given above. At least part of this material may, in fact, prove correct if further documentation becomes available. And then there are textual emendations in older editions, the status of which is often rather unclear. All these forms are therefore to be included – and properly commented – in the dictionary. And finally – lemmata, which are based on actual misread forms and are thus to be deleted from the present corpus, may in fact appear as clearly attested forms in the future.

## 9.5 Presentation of Material

### 9.5.1 Structure of Presentation

While the actual processing work is structured by inscriptions (see section 9.2.3, above), the presentation of the lexical material in the dictionary is generally structured by lexemes. Since Sabaic grammar does not really match the internal logic of alphabetical arrangement, ensuring usability proved rather tricky. While lexemes are operated in a standardized form for both internal and presentational purposes, this form is often difficult to reconstruct from an internal plural form<sup>24</sup>, or an irregular formed verbal stem<sup>25</sup>, and is thus not a reliable basis for arrangement or even search. Dictionaries for other languages with similar phenomena (such as Arabic or Gə‘əz) often opt for an arrangement by root. The latter, however, may also prove difficult to reconstruct and is thus an equally unreliable basis for search. We therefore decided to offer several different search options to ensure easy access to the material: lexemes proper, but also roots, strings of characters and translations.<sup>26</sup> All lexemes are complemented with a suggested translation, an automatically generated counter giving the number of attestations in the material processed thus far, a catalogue of attested morphological forms, a complete literary and etymological documentation and quotations of the particular lexeme in its syntactical and semantic context (Figure 9.1). For the time being, the presentation is exclusively in German. However, technical requirements for an eventual extension to other languages were taken into consideration.

<sup>24</sup> Such as *byt* “houses”, sg. *byt*, or *mšymt* “(agricultural) installations”, sg. *mšm*.

<sup>25</sup> Such as *htrgn* (infinitive) besides *thrg* (suffix conjugation) “to fight”.

<sup>26</sup> I.e., the German translations suggested in the dictionary.

**Suchergebnis für sb' 01** 131 Belege in 86 Texten

Übersetzung [Schrift verkleinern](#)

ausziehen, aufbrechen, zu Felde ziehen

Ältere Übersetzung [Aufklappen](#) [Schrift verkleinern](#)

(bellum) gerere  
CIH I, 408  
(eine Reise) unternehmen  
Sims 2000-24.Bsp. 8

Altsüdarabische  
Parallelen [Aufklappen](#) [Schrift verkleinern](#)

Qatabanisch  
sb'  
(einen Feldzug) unternehmen  
Lemus 1098-272

Etymologische  
Parallelen [Aufklappen](#) [Schrift verkleinern](#)

Arabisch  
masba' (Wz. sb') "wa-l-masba'u: t-tariqu fi l-ğabali 'masba' ist der Weg im Gebirge" Lisān III / 227  
sub'a (Wz. sb') "wa-'innaka la-turidu sub'atan 'ay turidu safaran ba'idan yuğayyiruka 'du

Formen [Aufklappen](#) [Schrift verkleinern](#)

SK  
3. m. sg. s[b']  
CIH 312/4<sup>2</sup>  
3. m. sg. sb'

Wendungen [Zurückklappen](#) [Schrift verkleinern](#)

sb' 'dy "ausziehen nach"  
b-kn sb' 'dy shrtm, Ry 538/14.-15., "als er nach Sahiratān zog"  
b-qt hwfy-hmy 'lmqh b-kn sb'y 'dy kbtm b-wrḥ ḡ-'lt ḡ-hrf 'b'mr bn wdd'l bn ḥzfrim ḡ-ḡmrn,  
Schm/Sir 133/4.-8., "dafür, daß 'Almaqah ihnen Wohlergehen geschenkt hat, als sie im

Figure 9.1: Result for lemma. The example is sb'

## 9.5.2 Accessible Material

### 9.5.2.1 Translation

An up-to-date translation is given. This is established on the basis of the complete epigraphic material. To enable retrograde search and thus usability, it is often enriched with synonyms. In ambiguous cases, all possible renderings are mentioned, including possible references to other lexemes.<sup>27</sup> Onomastic material is generally vocalized to facilitate reading. It should be kept in mind, however, that most of this vocalization is conventional.

<sup>27</sup> This is especially important for supposedly incorrect forms.

### 9.5.2.2 Existing Translations

A full catalogue of existing translations is being prepared.<sup>28</sup> This is particularly important since the meaning of many lexemes is still not sufficiently established in the literature. The whole range of possible interpretations should thus be made accessible to the user to reflect scientific discussion in the field. Part of the collected material stems from existing dictionaries such as Beeston et al., 1982, and glossaries to larger corpora such as Jamme, 1962, pp. 426–451. Unfortunately, the latter are not always comprehensive.<sup>29</sup> However, these sources in no way reflect the totality of existing interpretations. Further material is retrieved from editions (usually containing translations in context) and commentaries. In particular, translations referred to in commentaries to text editions are often dispersed and thus difficult to access.<sup>30</sup> Furthermore, existing translations of texts are thoroughly checked for their lexicographic content.<sup>31</sup> However, to allow a satisfactory workflow, translations have to be near to literal. Paraphrases, especially common for phrases considered as hendiadys, tend to be tricky. Nevertheless, these are included if they can be linked with certainty to particular lexemes.

The resulting catalogue is often surprisingly extensive. Since sufficient criteria for classifying variant forms as “identical” are missing, only verbatim quotes are considered as a single item. Catalogue entries may thus be rather close to each other. In exceptional cases such as *sb*’ “to go out”, approximately one hundred different translations have been collected, reaching from “(bellum) gerere” up to “zum Kriegszug aufbrechen”. While translations of common words normally started to center around a semantic nucleus at a comparatively early stage, sometimes back in the 19<sup>th</sup> century, translations of other lexemes can differ considerably over time. In particular, translations of forms with rather unspecific literal meanings, such as *sb*’, may also include a wide range of metaphorical renderings oriented towards context rather than literal meaning.

In the case of onomastics, all existing vocalizations are collected. As in the case of other lexemes, only verbatim quotes were subsumed under a single entry. However, this approach proved rather dysfunctional, given the huge amount of different

---

**28** Though some dictionaries (such as Biella, 1982 and Ricks, 1989) are commonly thought to be rather compilations of extant material than independent lexicographical work, this has never been done explicitly for the whole lexicon.

**29** This is e.g. the case in Beeston, 1976 (pp. 60–72), where only a “select glossary” is presented that is incomplete both in terms of lexemes and translations.

**30** Thus Ryckmans (1964, p. 91) has stated in his review of Jamme, 1962 “Ces suggestions ne manquent pas toujours d’intérêt en soi, mais outre qu’elles sont disséminées dans tout l’ouvrage, et par conséquent pratiquement perdues [...]”. It should be noted that commentaries often refer to lexemes not included in the commented text.

**31** This is mostly done while preparing the inscription that is currently processed. Translations from inscriptions which have not yet been processed are only sporadically included.

transliteration systems that entails numerous pseudovariants (e.g. Dât-Ḥimyam besides dhât Ḥimyam, the name of a female deity).

### 9.5.2.3 Etymological Parallels

Lemmata are enriched with etymological parallels from South Arabia and beyond. These are divided in non-Sabaic Ancient South Arabian, i.e. Qatabanic, Minaic and Ḥaḍramitic, and other Semitic languages. Ancient South Arabian parallels are catalogued according to Sabaic, i.e. a catalogue as complete as possible is intended. This includes material from both dictionaries and translations of actual texts, irrespective of correctness.

Since scientific literature (at least up to the first half of the 20<sup>th</sup> century) normally considered the different idioms as mere variants of a common Ancient South Arabian language, translations were often meant as applying to all respective languages. They can thus be considered as part of the catalogue of existing translations for Sabaic as well.<sup>32</sup> Etymological parallels from other Semitic languages are generally retrieved from the respective dictionaries. Cataloguing started with geographically adjacent languages (Arabic, Ethiopic and Modern South Arabian) and is by now far from complete, but is continuously being enriched.

### 9.5.2.4 Morphological Catalogue

An exhaustive morphological catalogue is created from the morphological tags. The section gives a fully referenced overview of attested forms,<sup>33</sup> which is arranged according to morphology.<sup>34</sup> All references are clickable and linked to quotations in context.

### 9.5.2.5 Examples in Context

The actual usage of each particular lexeme is illustrated by references within their contexts.<sup>35</sup> These are given both in transliteration of the Sabaic text and full German

---

<sup>32</sup> In some cases, translations in the secondary literature cannot be classified by language. These were normally taken as Sabaic.

<sup>33</sup> A complete catalogue of evidence is given in no other dictionary on the subject. This is, however, probably rather due to limited space and high printing costs than to editorial choices.

<sup>34</sup> Due to technical algorithms of the programme, incomplete forms from damaged passages split by brackets in the transliteration have to be given as separate entries. The catalogue of suffix conjugation forms of the 3<sup>rd</sup> person masculine singular of *ḥmr* “to grant” thus features by now a total of 8 different strings, including also entries like *ḥmr*], [*ḥmr*, *ḥm*], *ḥ*[*mr*, and so on besides complete *ḥmr*.

<sup>35</sup> Examples in contexts are given in Biella, 1982, though the collection is far from being exhaustive. The more concise presentation in Beeston et al., 1982 is devoid of such material.

translation. In order to demonstrate the different semantic and/or syntactical aspects of a word in different contexts, these quotations are structured by significant headlines that provide an overview of usages. For extensively attested lexemes,<sup>36</sup> however, even this classification will fail to guarantee a fairly clear presentation.

To ensure a consistent rendering and improve workflow, quoted passages are only translated once. Inscriptions are therefore split into paragraphs covering sufficiently comprehensible semantic units<sup>37</sup> and supplemented with a German translation. Renderings are kept as literal as possible. As all elements of texts are stored separately in the database, each chosen paragraph can be created using simple routines. The paragraphs thus created are subsequently allocated to their correct place in the structure of presentation of each single lexeme.

## 9.6 Results Reached Thus Far

The web presentation of the project, accessible under [<http://sabaweb.uni-jena.de>], was launched in 2016. At present (June 2018) it contains over 1,800 lemmata<sup>38</sup> (plus over 2,200 names) attested in around 900 inscriptions containing a total of 70,000 words. Choosing digital technology in preparing and presenting the dictionary certainly had substantial effects on the workflow. Huge amounts of diverging material such as inscriptions on the one hand and translations and etymologies on the other are easy to manage in the framework of a database. Lexicographical work can be structured according to internal criteria such as contexts and does not depend on external necessities such as alphabetical arrangement. The same applies to presentation, as lexemes can be published in any possible order without affecting usability.

**Acknowledgements:** The work presented in this present paper is, of course, not entirely my own. The team is made up of Norbert Nebes (project director), Anne Multhoff (responsible editor), Mariam Kilargiani (editor) and Heiko Werwick (software engineer). External cooperation partners are Ingo Kottsieper, Göttinger Akademie der Wissenschaften, and Peter Stein, Friedrich-Schiller-Universität Jena – Theologische Fakultät (minuscule inscriptions). The project started in 2012 and is scheduled for nine years. It is hosted at the Friedrich-Schiller-Universität Jena and generously funded by the Deutsche Forschungsgemeinschaft.

---

<sup>36</sup> Such as *hqny* “to dedicate”, showing more than 500 attestations. Since the project aims at a complete documentation of all attested material, it is almost impossible to keep the structure for such huge amounts of material visible.

<sup>37</sup> Though these may be complete sentences, the specific syntactical structure of Sabaic inscriptions often requires paragraphs on a sub-sentence level.

<sup>38</sup> This refers only to material publicly accessible. By now, basic information on over 4,000 different lexemes has been collected.

## Bibliography

- Arbach, M. (1993). *Le maḏābīen: Lexique - Onomastique et Grammaire d'une langue de l'Arabie méridionale préislamique. Tome I. Lexique maḏābīen. Comparé aux lexiques sabéens, qatabānite et ḥaḏramawtique* (PhD thesis). Aix-en-Provence: Université de Provence Aix-Marseille I.
- Arbach, M. (2001). Une photographie inédite de l'inscription Ir 13. *Raydān*, 7, 13–24.
- Arbach, M. & Schiettecatte, J. (2006). *Catalogue des pièces archéologiques & épigraphiques du Jawf au musée national de Ṣan'ā'.* Ṣan'ā': Centre français d'archéologie et de sciences sociales de Ṣan'ā'.
- Beeston, A.F.L. (1976). *Warfare in Ancient South Arabia (2nd. - 3rd. centuries AD)*. London: Luzac.
- Beeston, A.F.L., Ghul, M.A., Müller, W.W., & Ryckmans, J. (1982). *Sabaic Dictionary (English-French-Arabic)*. Louvain-la-Neuve: Peeters / Beyrouth: Librairie du Liban.
- Biella, J.C. (1982). *Dictionary of Old South Arabic. Sabaean Dialect*. Chico, CA: Scholars Press.
- CIH. *Corpus Inscriptionum Semiticarum. Pars quarta: inscriptiones ḥimyariticas et sabæas continens*. Paris: E Reipublicæ typographeo.
- al-'Iryānī, M.A. (1973). *Fī tāriḥ al-Yaman. Ṣarḥ wa-ta'liq 'alā nuqūš lam tunṣar. 34 naqṣan min maḡmū'at al-Qāḏī 'Alī 'Abdallāh al-Kuhālī*. al-Qāhira.
- Jamme, A. (1962). *Sabean Inscriptions from Maḥram Bilqīs (Mārib)*. Baltimore, MD: John Hopkins Press.
- Jamme, A. (1985). *Miscellanées d'ancien arabe XIV*. Washington [privately printed].
- Lundin, A.G. (1987). Sabaean Dictionary. Some Lexical Notes. In C. Robin & M. Bāfaqīh (Eds.), *Ṣayḥadica. Recherches sur les Inscriptions de l'Arabie Préislamique Offertes par ses Collègues au Professeur A.F.L. Beeston* (pp. 49–56). Paris: Geuthner.
- Maraqten, M. (2014). *Altsüdarabische Texte auf Holzstäbchen. Epigraphische und kulturhistorische Untersuchungen*. Würzburg: Ergon.
- Multhoff, A. (2011). *Die Verbalstambildung im Sabäischen* (PhD thesis). Jena: Friedrich-Schiller-Universität Jena.
- Prioleta, A. (2013). *Inscriptions from the southern highlands of Yemen. The epigraphic collections of the museums of Baynūn and Dhamār* (Arabia Antica 8). Roma: L'«Erma» di Bretschneider.
- RES. *Répertoire d'Épigraphie Sémitique. Publié par la commission du Corpus Inscriptionum Semiticarum*. Paris: Imprimerie nationale.
- Ricks, S.D. (1989). *Lexicon of Inscriptional Qatabanian*. Roma: Editrice Pontificio Istituto Biblico.
- Robin, Ch.J. (2013). À propos de *Ymnt* et *Ymn*: “nord” et “sud”, “droite” et “gauche”, dans les inscriptions de l'Arabie antique. In Fr. Briquel-Chatonnet, C. Fauveaud, & I. Gajda (Eds.), *Entre carthage et l'Arabie heureuse. Mélanges offerts à François Bron* (pp. 119–140). Paris: De Boccard.
- Ryckmans, J. (1964). Review of A. Jamme, *Sabaean Inscriptions from Maḥram Bilqīs (Mārib)*, with foreword by Wendell Phillips. Baltimore, The Johns Hopkins Press, 1962 (...). *Bibliotheca Orientalis*, 21, 90–94.
- Schlobies, H. (1936). Neue Dokumente zur altsüdarabischen Epigraphik. *Orientalia*, 5, 57–63.
- Sima, A. (2000). *Tiere, Pflanzen, Steine und Metalle in den altsüdarabischen Inschriften. Eine lexikalische und realienkundliche Untersuchung*. Wiesbaden: Harrassowitz.
- Stein, P. (2002). Schreibfehler im Sabäischen am Beispiel der mittelsabäischen Widmungsschriften. *Le Muséon*, 115, 423–467.
- Stein, P. (2003). *Untersuchungen zur Phonologie und Morphologie des Sabäischen*. Rahden/Westf.: Marie Leidorf.

- Stein, P. (2010). *Die altsüdarabischen Minuskelinschriften auf Holzstäbchen aus der Bayerischen Staatsbibliothek in München. Bd. 1. Die Inschriften der mittel- und spätsabäischen Periode.* Tübingen/Berlin: Wasmuth.
- Stein, P. (2011). Ancient South Arabian. In S. Weninger (Ed., in collaboration with G. Khan, M.P. Streck, & J.C.E. Watson), *The Semitic Languages. An International Handbook* (pp. 1042–1073). Berlin/Boston: De Gruyter.



Ronald Ruzicka

## 10 KALAM: A Word Analyzer for Sabaic

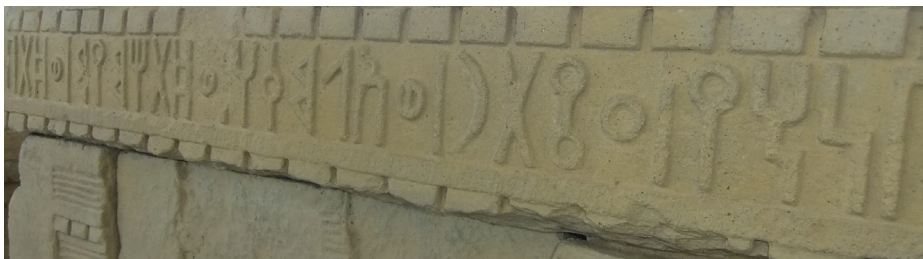
**Abstract:** The word analyzer KALAM for Sabaic is a tool for the automatic detection of morphological attributes of a Sabaic word, like stem, conjugation, case and person. It has been developed as part of a Masters thesis in Arabic Studies. Connected to a computer-based dictionary, it also provides the translation, including prefixes and postfixes, like possessive pronouns and particles. New research work has connected the new system “KALAM reloaded” to online dictionaries like the Sabaic Online Dictionary, and is now extended to Minaic, Qatabanic and Ḥaḍramitic, too. The final aim is the automatic translation of sentences of Ancient South Arabian languages. The development of the project will be aided by using the newly digitized texts of the Glaser collection at the Austrian Academy of Sciences and building up annotated trees in a database in an iterative process, improving the algorithms.

**Keywords:** Sabaic, Ancient South Arabian, word analyzer, translation, squeezes

### 10.1 An Automatic Word Analyzer for Languages Epigraphically Attested

The word analyzer KALAM for Sabaic words has been developed within the framework of a Masters thesis (Ruzicka, 2016) in Arabic Studies at the University of Vienna’s Institute for Oriental Studies with the main emphasis on South Arabia.

The objective was to find the roots of words or word-phrases, i.e. everything between two word dividers in Ancient South Arabian languages. Figure 10.1 shows a sample Sabaic text from Ethiopia.



**Figure 10.1:** Altar inscription, Wukro, Ethiopia (photo by R. Ruzicka)

---

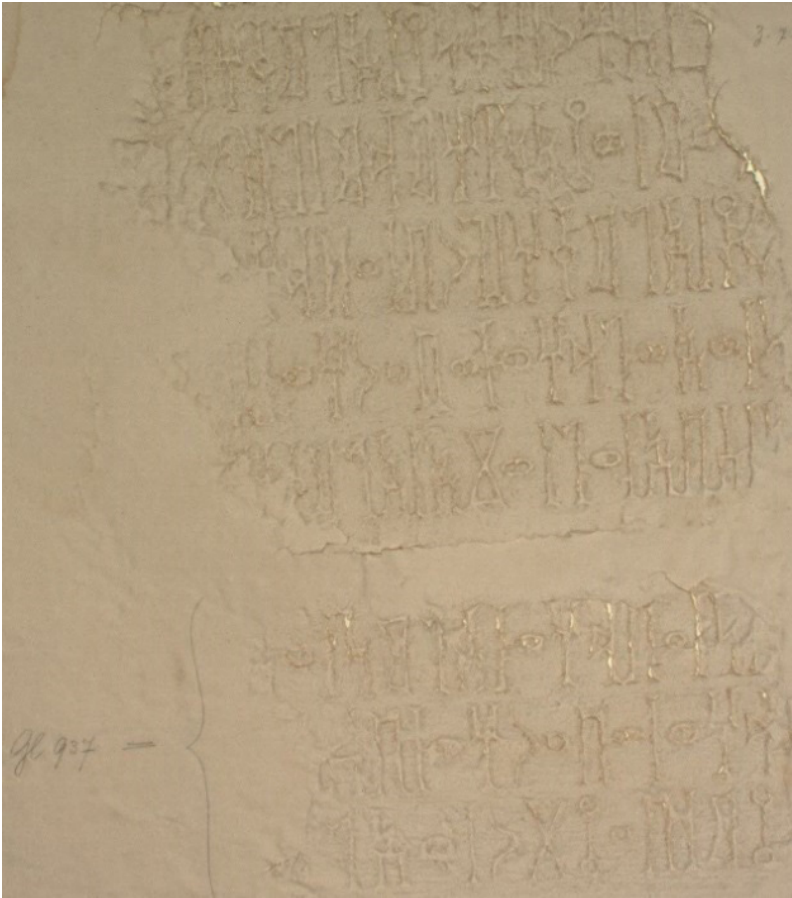
Ronald Ruzicka, Simutech and Austrian Academy of Sciences, Vienna



© 2018 Ronald Ruzicka

This work is licensed under the Creative Commons Attribution 4.0 International license (CC BY 4.0)

The word analyzer supported the decipherment of all the theoretically possible meanings and readings of a word-phrase or a word-phrase with *lacunae*. The squeezes of the Glaser collection (Höfner, 1944) at the Austrian Academy of Sciences (Figure 10.2), which are being digitized by a project in which the author is also participating, contain a significant volume of unreadable text, meaning that sometimes only small portions of words are available. Within the digitizing project<sup>1</sup> words or word-phrases are tagged for computational usage. This means that names and ordinary words are separate and that, in due course, all words will be classified according to their grammatical features.



**Figure 10.2:** Squeeze sample collected by Eduard Glaser

<sup>1</sup> Glaser project's web-application [<http://glaser.acdh.oeaw.ac.at/>].

This tagging is intended to be carried out automatically by a computer. This is an important desideratum given that, in the DASI project,<sup>2</sup> the grammatical features of a word can at present be assigned only manually (Avanzini et al., 2015).

The idea was to develop a computer service, which takes a word-phrase in an Ancient South Arabian language as input and delivers the grammatical analysis as output: all possibilities for root, stem, person, number, gender, time, and origin, i.e. Sabaic, Minaic, Qatabanic, Ḥaḍramitic (see Avanzini, 2009 for the background of the classification of the languages). It should also separate prefixes, suffixes, conjunctions, and so on.

## 10.2 Requirements of the Word Analyzer for Sabaic

In this section, the operations that the word analyzer should be able to perform will be explained with examples taken from the Ancient South Arabian languages. Let us consider the word *stṣr*<sup>3</sup>. The root of this verb is:

*nṣr*  
*ST-stem, 3rd masculine singular, 3rd feminine plural*  
*Middle and Late Sabaic*  
 to call somebody for support

Given that Ancient South Arabian displays weak radicals in some words, however, the root could conceivably be construed as (Multhoff, 2012):

*\*ṣwr, which is a hypothetical root*  
*ST-stem*  
*\*ṣyr*  
*ST-stem*

The pattern of words where the roots cannot be found easily, or where the result is not unique, in a form like ABCD, could be:

*stem ABC + suffix D ?*  
*prefix A + stem BCD*  
*prefix A + stem BCC + suffix D*

The word analyzer should provide all possible variations of grammatical features that fit ABCD in a certain case.

<sup>2</sup> DASI project [<http://dasi.cnr.it/>].

<sup>3</sup> In the whole project, we only deal with transcribed characters. There is an online keyboard available to enter diacritic characters.

The base of the KALAM service can be found in the most recent book on Sabaic grammar, the *Lehrbuch der sabäischen Sprache* by Peter Stein (Stein, 2013), which contains the whole paradigm for Sabaic – as far as it is known today.

The algorithm for the computer service is obtained through a synthetical approach. The program reproduces the scholar's reasoning. It first searches for the root, then synthesizes all possible epigraphically attested forms based on this root. Then it compares these forms to the input word-phrase.

Translation programs for Semitic languages were already available, such as the *Stanford Log-linear Part-Of-Speech Tagger*<sup>4</sup> and the *Arabic Language Analyzer with Lemma Extraction and Rich Tagset* (Aliwy, 2012), this last being a very complex but powerful translator for Arabic.

Nevertheless, all of the translation methods display disadvantages as word analyzers. Indeed, focusing on living languages, all of them are based on very large corpora of digital text available in the web, whereas the ancient languages of the Arabian Peninsula are only fragmentarily attested by inscriptions. They have further problems in cases where no vowels are available, or where rules changed over the course of time.<sup>5</sup> Furthermore, some of them cannot deal with the assimilation of consonants or with weak radicals.

### 10.3 Functioning of the Word Analyzer

The aforementioned considerations prompted the creation of KALAM from scratch as a word analyzer for the Sabaic language.

KALAM uses three main categories with which to analyze the grammar:

- the so-called *Situation*, which consists of the *TimePeriod* and the *Locality*, for instance Middle and Late Sabaic;
- the *Rule*, together with the *rule-situation*, for instance verbs with suffix conjugation in a certain time frame, leading to
- a *Term*, which describes how, for example, the first common singular form is created.

Thus, looking at the 3 radicals (or 4 in some cases) we have possible prefixes, infixes and suffixes. Metathesis, gemination and weak radicals could occur as well. Moreover, all special cases of words, which do not fit in the ordinary schemes, must be taken into account.

Here are samples of the configuration list, for instance a verb in prefix conjugation, first person common singular, which carries the prefix ʾ (aleph) (symbolized by an “a”

<sup>4</sup> [<https://nlp.stanford.edu/software/tagger.shtml>].

<sup>5</sup> Sabaic and its brother languages Minaic, Qatabanic, Ḥaḍramitic have been written from the 9<sup>th</sup> century BCE up to the 6<sup>th</sup> century CE.

here), and some special rules for weak radicals. Some weaknesses are unavoidable (writing 0), some optional (writing 0 and 1).

*v\_PKO\_1cs*  
*a null null null 0 0 n\*\*=011 w\*\*=011 \*w\*=101,111 \*y\*=101,111 \*..=110 ydc=011*

The following one is another sample for a noun in status constructus and dual form:

*n\_SC\_mdn*  
*null null null y 0 0 n\*\*=011,111 w\*\*=011,111 \*w\*=101,111 \*y\*=101,111*

In this way, about 670 rules have been created which build up the whole grammar of Sabaic, including special cases such as:

- cardinal and ordinal numbers,
- parts,
- personal/demonstrative pronouns,
- suffix/object pronouns,
- particles,
- conjunctions,
- *nomen loci*.

### 10.3.1 Using KALAM

Using KALAM is quite simple (Figure 10.3). Entering a word phrase, for instance “*ybny*”, and selecting the mode and the language (only “Sabaic” in this case), a possible root, the stem, a short and a long description of the grammar term are provided.

Root	Stem	Term	Location	Time	Prefix	Suffix	Translation KALAM	Information
bny	O1	sabv_PKO_3ms	Saba'	middle+late-sabaic,old-sabaic			build, construct,building, edifice,(act of) building, construction	verb; prefix conjugation; 3; masculine singular
bny	O1	sabv_PKO_3md	Saba'	old-sabaic			build, construct,building, edifice,(act of) building, construction	verb; prefix conjugation; 3; masculine dual
bny	O2	sabv_O2PKO_3ms	Saba'	middle+late-sabaic,old-sabaic			build, construct,building, edifice,(act of) building, construction	verb; prefix conjugation; 3; masculine singular
bny	O2	sabv_O2PKO_3md	Saba'	old-sabaic			build, construct,building, edifice,(act of) building, construction	verb; prefix conjugation; 3; masculine dual
byn	O1	sabv_PKO_3md	Saba'	middle+late-sabaic			between; among; in both of (two things);intervene, separate (boundary);remove (punishment);surrounding are (of town); surrounding loculi (of tomb)	verb; prefix conjugation; 3; masculine dual

Figure 10.3: Analyzing *ybny* using KALAM

If we enter the word “stšr” we will get any possible root – even the hypothetical ones listed above (see 10.2).

Forms beginning with the letter s are also displayed (Figure 10.4):

stšr

Root	Stem	Term	Location	Time	Prefix	Suffix	Translation KALAM	Information
nšr	O1	sabv_PK0_2ms	Saba'	middle+late-sabaic	sabpart_s		at,near to	verb; prefix conjugation; 2. masculin singular
nšr	O1	sabv_PK0_3fs	Saba'	middle+late-sabaic	sabpart_s		at,near to	verb; prefix conjugation; 3. feminin singular
nšr	O1	sabv_PK0_3fp	Saba'	middle+late-sabaic	sabpart_s		at,near to	verb; prefix conjugation; 3. feminin plural
nšr	ST	sabv_STSK_3ms	Saba'	middle+late-sabaic				verb; suffix conjugation; 3. masculin singular
nšr	ST	sabv_STSK_3fp	Saba'	middle+late-sabaic				verb; suffix conjugation; 3. feminin plural
nšr	TI	sabv_T1IMPk_2ms	Saba'	sabaic	sabpart_s		at,near to	verb; imperative; 2. masculin singular

Figure 10.4: Analyzing stšr using KALAM

If the hypothetical roots are not desired, an inbuilt word list can be used – a small dictionary that not only reduces the roots to already existing ones, but also provides a translation. If “use dictionary” is checked, only really existing words are displayed (Figure 10.5).

As one of the goals of the project is to support students as well as established scholars, the system allows a root, like QTL, to be entered, and the command “synthesize” will then display all grammatical forms the paradigm knows.

, e d q h h

š š š t t z

enter  ASA  English word  
 use dictionary  entire dictionary  use SabaWeb  4 radicals, too  
 sabaic  minaic  qatabanic  hadramitic

mhfd

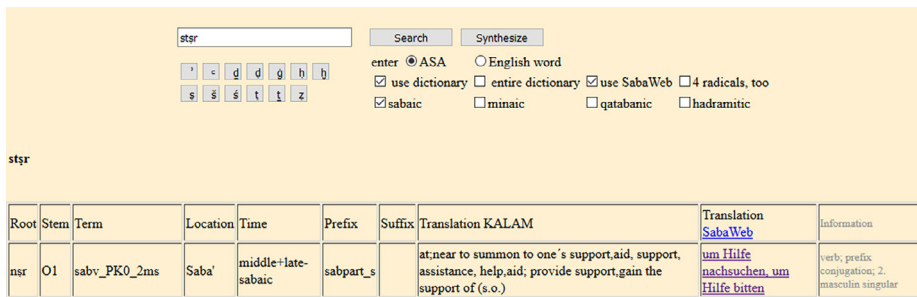
Root	Stem	Term	Location	Time	Prefix	Suffix	Translation KALAM	Information
hfd	O1	minn_LOO1SI_cs	Macin	minaic			tower; projecting element (of wall)	nomen loci; status indeterminatus; communis singular
hfd	O2	minn_PZO2SI_cs	Macin	minaic			tower; projecting element (of wall)	passive/active participle; communis singular
hfd	O2	minn_LOO2SI_cs	Macin	minaic			tower; projecting element (of wall)	nomen loci; status indeterminatus; communis singular
hfd	O1	sabn_LOO1SC_ms	Saba'	sabaic			tower; projecting element (of wall)	nomen loci; status constructus; masculin singular
hfd	O2	sabn_PZO2SC_ms	Saba'	sabaic			tower; projecting element (of wall)	passive/active participle; masculin singular
hfd	O2	sabn_LOO2SC_ms	Saba'	sabaic			tower; projecting element (of wall)	nomen loci; status constructus; masculin singular

Figure 10.5: Using the dictionary. If one enters “mhfd” selecting “Sabaic”, “Minaic” and “use dictionary” one obtains “tower”. Similarly, entering “tower” and selecting “English word” produces “mhfd”

## 10.4 Future Perspectives

KALAM can be used freely online<sup>6</sup> and presently supports all four Ancient South Arabian languages (Minaic, Qatabanic and Ḥaḍramitic in addition to Sabaic). The user can search for word-phrases with single missing characters, entering a question mark instead of a character. Connections to other online tools for the Ancient South Arabian languages, like the Sabaic Online Dictionary (University of Jena) or DASI (University of Pisa) are provided.

For DASI, some pre-work has been done. For the Sabaic Online Dictionary (hereafter, SabaWeb), which provides all known occurrences and state-of-the-art translations of the words in Sabaic, we have a full integration reached (Figures 10.6, 10.7).



The screenshot shows the KALAM search interface. At the top, there is a search input field containing 'stṣr' and buttons for 'Search' and 'Synthesize'. Below the input field are two rows of character selection buttons: the first row contains 's', 'c', 'd', 'd', 'g', 'h', 'h' and the second row contains 's', 'g', 'š', 't', 'z'. To the right of the input field are radio buttons for 'ASA' (selected) and 'English word', and checkboxes for 'use dictionary' (checked), 'entire dictionary', 'use SabaWeb' (checked), and '4 radicals, too'. Below these are checkboxes for 'sabaic' (checked), 'minaic', 'qatabanic', and 'hadramitic'. The search results are displayed in a table with the following columns: Root, Stem, Term, Location, Time, Prefix, Suffix, Translation KALAM, Translation SabaWeb, and Information.

Root	Stem	Term	Location	Time	Prefix	Suffix	Translation KALAM	Translation SabaWeb	Information
nsr	O1	sabv_PK0_2ms	Saba'	middle+late-sabaic	sabpart_s		at.near to summon to one's support,aid, support, assistance, help,aid; provide support,gain the support of (s.o.)	<a href="#">um Hilfe nachsuchen</a> , <a href="#">um Hilfe bitten</a>	verb, prefix conjugation; 2, masculin singular

**Figure 10.6:** Connecting KALAM and SabaWeb. We take again “stṣr”, “use dictionary”, “Sabaic only” and select “SabaWeb”. Clicking on the SabaWeb link, all additional information is shown



The screenshot shows the SabaWeb search results page. At the top, there is a header with the logo of the Deutsche Forschungsgemeinschaft (DFG) and the Friedrich-Schiller-Universität Jena. Below the header, there is a navigation menu with links for 'Projekt', 'Suche', 'Corpus', 'Statistik', 'Bibliographie', 'Kontakt', 'Andere Projekte', and 'Hilfe'. The main content area is titled 'Suche' and shows the search results for 'nsr ST'. It includes a citation 'Zitierform sabaweb.uni-jena.de [Zugriff am 22.08.2016]', the search result 'Suchergebnis für nsr ST' with '3 Belege in 3 Texten', and a table of results with columns for 'Übersetzung' and 'Ältere Übersetzung'. The 'Übersetzung' column shows 'um Hilfe nachsuchen, um Hilfe bitten' and the 'Ältere Übersetzung' column shows '(Bitte um) Unterstützung richten' with references to Nebes 2005, 350 and appeler au secours Calvet/Robin 1997, 144.

**Figure 10.7:** Result page from SabaWeb

<sup>6</sup> KALAM [http://kalam.ruzicka.net].

SabaWeb is linked to KALAM for two purposes:

- it detects already existing roots;
- it delivers all the translation features described above.

A new project at the Austrian Academy of Sciences starting in 2018 will try new methods of scanning the squeezes of the Glaser collection (Figure 10.2) and will provide not only digitized, but fully tagged texts. KALAM will be “reloaded”, too, and get new features. The aim is to advance from word-phrases to parts of sentences by including the automatic building of syntax trees. The currently manually tagged texts of the Glaser collection will then provide a small corpus with which the algorithms for automatic tagging in KALAM can be improved and parametrized.

The research question of this project is: is it possible to fill in missing parts of text automatically? Using word analysis, automatic tagging and “sentence formulars”, as they appear for instance in dedication and building inscriptions, it should be possible. The comparison of syntax trees will be used within the corpus. At least the question is: to which degree can we support semi-automatic filling of *lacunae* in the text?

## Bibliography

- Aliwy, A. (2012). Arabic Language Analyzer with Lemma Extraction and Rich Tagset. In H. Isahara & K. Kanzaki (Eds.), *Advances in Natural Language Processing: Proceedings of the 8th International Conference on NLP, JapTAL 2012, Kanazawa, Japan, October 22-24, 2012* (pp. 168–179). Springer-Verlag. doi: 10.1007/978-3-642-33983-7\_17
- Avanzini, A. (2009). Origin and classification of the Ancient South Arabian languages. *Journal of Semitic Studies*, 54, 205–220.
- Avanzini, A., De Santis, A., Gallo, M., Marotta, D., & Rossi I. (2015). Computational Lexicography and Digital Epigraphy: Building digital lexica of fragmentary attested languages in the Project DASI. In G. Guidi, R. Scopigno, J.C. Torres, & H. Graf (Eds.), *2015 Digital Heritage International Congress* (pp. 405–408). New York: IEEE. doi: 10.1109/DigitalHeritage.2015.7419535
- Höfner, M. (1944). *Die Sammlung Eduard Glaser* (Akademie der Wissenschaften in Wien, Philosophisch-historische Klasse, Sitzungsberichte 222.5). Brunn-München-Wien: Rudolf M. Rohrer.
- Multhoff, A. (2012). *Die Verbalstambildung im Sabäischen* (PhD thesis). Jena: Friedrich-Schiller-Universität Jena.
- Ruzicka, R. (2016). *KALAM – Wortanalyse für Sabäisch* (MA thesis). Wien: Universität Wien.
- Stein, P. (2013). *Lehrbuch der sabäischen Sprache. Teil 1: Grammatik*. Wiesbaden: Harrasowitz Verlag.



Jamie Novotny and Karen Radner

# 11 Official Inscriptions of the Middle East in Antiquity: Online Text Corpora and Map Interface

**Abstract:** The LMU Munich-based Official Inscriptions of the Middle East in Antiquity (OIMEA) project is one of the two principal, digital text corpora of the Munich Open-access Cuneiform Corpus Initiative (MOCCI), which is a freely accessible digital humanities umbrella project established by Karen Radner and Jamie Novotny in the fall of 2015. This international project – which includes research partners in Philadelphia, Barcelona, and Rome – aims to edit all available official inscriptions of ancient Middle Eastern polities, recorded in the cuneiform script and contemporary writing systems, in a freely accessible, fully lemmatized (lexical and grammatical data tagging), and completely searchable format via the Open Richly Annotated Cuneiform Corpus (Oracc) project. In addition, OIMEA plans to make geo-referenced text editions available through its Ancient Records of Middle Eastern Polities (ARMEP) map interface, which is developed in collaboration with LMU’s Center for Digital Humanities.

**Keywords:** Assyria, Babylonia, geo-referencing, lemmatization, map interface

## 11.1 Introduction

In September 2015, the present authors founded Official Inscriptions of the Middle East in Antiquity (OIMEA) as part of the newly established Chair for the Ancient History of the Near and Middle East at LMU Munich. The aim was to widely disseminate, facilitate, and promote the active use and understanding of royally-composed texts of ancient Middle Eastern polities in academia and beyond, and to begin creating new and innovative ways for users to access the important and varied contents of these geo-referenced and linguistically-annotated (lemmatized) ancient records.<sup>1</sup>

---

<sup>1</sup> MOCCI [<http://www.en.ag.geschichte.uni-muenchen.de/research/mocci/index.html>] and OIMEA [<http://oracc.museum.upenn.edu/oimea/index.html>]. MOCCI’s other digital text corpus, Archival Texts of the Middle East in Antiquity (ATMEA), which presently consists only of State Archives of Assyria online – SAAo [<http://oracc.museum.upenn.edu/saao/index.html>], is not discussed here. A third LMU cuneiform text corpus, Electronic Babylonian Literature (eBL), will be developed by Enrique Jiménez starting in 2018.

---

Jamie Novotny, Karen Radner, Ludwig-Maximilians-Universität, München



© 2018 Jamie Novotny and Karen Radner

This work is licensed under the Creative Commons Attribution 4.0 International license (CC BY 4.0)

Given our research interests and past project affiliations, our initial focus was naturally the Akkadian and Sumerian inscriptions of Mesopotamia, primarily the self-aggrandizing texts of the kings of Assyria and Babylonia from ca. 1157–539 BCE. Thus, we initiated the following three projects: Royal Inscriptions of Assyria online,<sup>2</sup> Royal Inscriptions of Babylonia online,<sup>3</sup> and Inscriptions of Suhu online.<sup>4</sup> In late 2016, in order to expand the dataset beyond Mesopotamia proper, work began on two other inscription-focused projects: the Electronic Corpus of Urartian Texts – eCUT, which contains cuneiform texts written in the Urartian language from Eastern Turkey, Armenia and Northwestern Iran; and Achaemenid Royal Inscriptions online – ARIO, which includes cuneiform texts written not only in Old Persian, but also in Elamite and Akkadian, chiefly from Iran.<sup>5</sup>

All five projects<sup>6</sup> include on the one hand, informational portal pages with details about the rulers in whose names these texts are written, their polities and the texts themselves; and on the other hand, the linguistically-annotated (lemmatized) editions with translations into English and, depending on the heritage data, also other European languages (German for RIBo), as well as the glossaries created from these editions. The text corpora are either (retro)digitized or newly created using software developed by Steve Tinney (Philadelphia) and are hosted on the Open Richly Annotated Cuneiform Corpus (Oracc) platform.

Lastly, in December 2016, the present authors, together with staff of the LMU's Center for Digital Humanities directed by Christian Riepl and Stephan Lücke, began developing a map interface designed to display places where ancient texts were discovered or that are mentioned in those ancient sources. The main purpose of this interface is to allow users access to Oracc-hosted texts directly from the map. Ancient Records of Middle Eastern Polities 1.0<sup>7</sup> was made public in December 2017. ARMEP's gazetteer feature, which displays cities mentioned in ancient sources, has not yet been implemented and will be part of version 2.0, which is to be developed in 2018.

---

<sup>2</sup> RIAo [<http://oracc.museum.upenn.edu/riao/index.html>].

<sup>3</sup> RIBo [<http://oracc.museum.upenn.edu/ribo/index.html>].

<sup>4</sup> Suhu [<http://oracc.museum.upenn.edu/suhu/index.html>].

<sup>5</sup> No URLs are provided for ARIO and eCUT as neither are yet publically accessible. Both are planned for release in 2018.

<sup>6</sup> The count is six, when one includes the University of Pennsylvania-based Royal Inscriptions of the Neo-Assyrian Period (RINAP) Project (directed by Professor Grant Frame). Radner has been a member of that project's editorial board since its inception in 2008 and Novotny has been a principal content contributor to both its printed books and its freely accessible online content since 2009; he is currently preparing new, online editions of the royal inscriptions of Assyria's last great king Ashurbanipal and his lesser-known successors (668–612 BCE).

<sup>7</sup> ARMEP [<https://www.armep.gwi.uni-muenchen.de>].

This paper will briefly discuss OIMEA and ARMEP, as well as address some methodological problems and technical issues in their creation and the future prospects of these two projects.

## 11.2 Overview of OIMEA and Its Sub-Projects

As is obvious from its name, the scope of OIMEA is official inscriptions, primarily from Middle Eastern polities of the first millennium BCE. The idea was inspired by the now-defunct, Toronto-based Royal Inscriptions of Mesopotamia – RIM Project. This project had attempted to edit, in a single place, royal inscriptions written in the Akkadian and Sumerian languages. Our Oracc-hosted umbrella project and search tool, which intends to go beyond the scope of the RIM Project, currently comprises six projects:

- Corpus of Kassite Sumerian Texts,<sup>8</sup>
- Electronic Text Corpus of Sumerian Royal Inscriptions,<sup>9</sup>
- Royal Inscriptions of Assyria online (LMU Munich),
- Royal Inscriptions of Babylonia online (LMU Munich),
- Royal Inscriptions of the Neo-Assyrian Period online,<sup>10</sup> and
- Suhu (LMU Munich)<sup>11</sup>

The presently available texts on OIMEA are all written in the Akkadian and Sumerian languages and in cuneiform. Starting in 2018, the project will include corpora of texts written in other languages; for example, monolingual Old Persian and trilingual Persian, Elamite and Akkadian inscriptions will be included on ARIO,<sup>12</sup> and monolingual Urartian and bilingual Urartian and Assyrian inscriptions will be accessible via eCUT.<sup>13</sup>

---

**8** CKST [<http://oracc.museum.upenn.edu/ckst/index.html>], University of California Berkeley.

**9** ETCRSRI [<http://oracc.museum.upenn.edu/etscri/index.html>], Eötvös Loránd University Budapest.

**10** RINAPo [<http://oracc.museum.upenn.edu/rinap/index.html>], University of Pennsylvania and LMU Munich.

**11** Suhu contains retro-digitized and lemmatized editions of the officially commissioned texts of the extant, first-millennium-BCE inscriptions of the rulers of Suhu; these texts were published in Frame (1995, pp. 275–331). The open-access transliterations and translations were lemmatized and updated by Alexa Bartelmus.

**12** The contents of ARIO are based primarily on Schmitt, 2009, as well as data provided by Matt Stolper (Chicago) from his now-defunct Achaemenid Royal Inscriptions project [<https://oi.uchicago.edu/research/projects/achaemenid-royal-inscriptions-project>]. ARIO is currently managed by Henry Heitmann-Gordon.

**13** The contents of eCUT are based on Salvini (2008–12). Birgit Christiansen is currently retro-digitizing, updating, and lemmatizing, as well as translating into English, that corpus of inscriptions.

The OIMEA hub on Oracc primarily serves as a multi-project search engine that enables anyone interested in the genre of official inscriptions to simultaneously search the translations, transliterations, catalogues, and portal pages of every available project on which ancient inscriptions are edited.<sup>14</sup> As an informational and search hub, OIMEA strives to make the vast and varied corpora of inscriptions easily and freely accessible to every scholar, student, and interested member of the general public. Moreover, it enables its users to efficiently search that rich genre of ancient records, allowing them, for instance, to perform searches both on the transliterations and on the translations.<sup>15</sup>

To give the readers of this volume a better idea of some of the content produced by OIMEA, two of the LMU-based projects, RIAo and RIBo, will be briefly described here.

### 11.2.1 Royal Inscriptions of Assyria Online

RIAo is intended to present up-to-date editions of officially commissioned texts of the rulers of Assur and later Assyria from the end of the third millennium BCE to the fall of Nineveh in 612 BCE; it also includes numerous informational portal pages that provide the historical and cultural contexts of these important ancient sources.<sup>16</sup> The project started in September 2015 and Phase 1 of the website was made available to the public in early 2016, when the project's initial goal – the retro-digitization of 866 Assyrian inscriptions published in three discipline-standard monographs (Grayson, 1987, 1991, 1996) – had been realized; these texts date from the end of the third millennium BCE to 745 BCE. Phase 2 of RIAo was completed in February 2018. That stage included the full lemmatization (lexical and grammatical data tagging) of every available text, as well as the completion of glossaries of the Akkadian words and proper names (gods, people, places, and temples) appearing in those 866 inscriptions and the writing of numerous informational pages on Assyria's many rulers; this work was principally carried out by Nathan Morello. The project is now entering Phase 3, which will consist of:

---

**14** A new “pager” interface is being developed for Oracc and its inspiration is based on the conceptual design and easy-to-search functionality of OIMEA. Oracc's “Neo” interface will allow users to easily and efficiently navigate and search all of the publically available texts on Oracc.

**15** For example, if one searches for “lion”, 91 matches are found in inscriptions from Early Dynastic times to the Neo-Babylonian Period; and if one searches for “scribe”, 209 matches are displayed for Sumerian and Akkadian texts written in the third, second, and first millennia BCE.

**16** For further details, see [<http://oracc.museum.upenn.edu/riao/abouttheproject/index.html>]. RIAo's focus is restricted to texts written in the Akkadian language, in the cuneiform script. It does not include information or access to documents pertaining to Modern Assyrians. For such a project, see the Modern Assyrian Research Archive Project [<http://assyrianarchive.org/database/home/>].

- incorporating the material published by the Royal Inscriptions of the Neo-Assyrian Period (RINAP) Project (directed by Professor Grant Frame and based at the University of Pennsylvania in Philadelphia) into the dataset;
- adding score transliterations of inscriptions known from more than one exemplar;
- supplementing composite editions with individual object transliterations when these are accessible for study (in the form of photographs, hand-drawn facsimiles, or in a museum or private collection);
- writing additional portal pages on rulers and their inscriptions; and
- preparing a comprehensive bibliography of Assyrian royal inscriptions.

When RIAo is finished, it will contain fully lemmatized and completely searchable editions of the approximately 1,800 Assyrian inscriptions.<sup>17</sup> By 2020, the complete corpus of Assyrian inscriptions will be easily accessible to scholars, students, and the general public. Anyone interested in Assyrian culture, history, language, religion, and texts will be able to efficiently search any Akkadian and Sumerian words appearing in the inscriptions and any English word used in the translations.

Content undergoes strict scientific control. Unlike community-built sites such as Wikipedia, Wikidata, Pelagios, and Pleiades, RIAo's contents cannot be created or edited by anyone. This is the sole responsibility of the core OIMEA team (presently Morello and Novotny), with input from OIMEA's international editorial and advisory boards. The present authors, as the directors of MOCCI, assume content and editorial oversight of the project. We do welcome/encourage feedback from our community of users.<sup>18</sup>

### 11.2.2 Royal Inscriptions of Babylonia Online

RIBo, which was also founded by the authors in September 2015, intends to publish in a single place fully searchable, lemmatized editions of all of the known Akkadian and Sumerian royal inscriptions from Babylonia that were composed between 1157 and 64 BCE, together with informational portal pages and complete glossaries of Akkadian and Sumerian words and the names of gods, people, places, and temples. The scope, when compared to RIAo, is much smaller. By the time RIBo is completed,

---

<sup>17</sup> Presently published in Grayson, 1987, 1991, 1996; Leichty, 2011; Tadmor & Yamada, 2011; Grayson & Novotny, 2012, 2014; Novotny & Jeffers, 2018, 2019; Frame, 2019.

<sup>18</sup> However, neither the scholarly community nor general public make much use of the possibility to contact the project by email (via the "About the Project" pages).

which is anticipated to be in 2022, that open-access project will contain about 400 inscriptions.<sup>19</sup>

Unlike RIAo, which comprises a single text corpus, the contents of RIBo are divided into several sub-corpora, generally grouped by “dynasty” or period.<sup>20</sup> However, all of these sub-corpora will be accessible from one interface.<sup>21</sup>

Phase 1 was first made public in early 2016 and was completed in early 2018. The content created includes:

- lemmatized editions of the inscriptions published in Frame 1995 and Da Riva 2013, as well as the famous “Cyrus Cylinder” and the “Antiochus (Borsippa) Cylinder”; and
- numerous informational portal pages on Babylonian rulers and their inscriptions, as well as on the various Babylonian King Lists.

During Phase 2, scheduled to run from 2018–2022, RIBo will produce fully lemmatized and searchable editions of the complete corpus of royal inscriptions of the six rulers of the Neo-Babylonian Empire (625–539 BCE): Nabopolassar, Nebuchadnezzar II, Amel-Marduk, Neriglissar, Labaši-Marduk, and Nabonidus. The transliterations, translations (English, as well as German), and glossaries (Akkadian, Sumerian, and proper names) will be fully searchable.

Eventually, RIBo will also include official inscriptions from the second millennium BCE, namely of the First Dynasty of Babylon, and the Kassite Period, but these corpora are being assembled/prepared by project partners in Munich and Philadelphia.<sup>22</sup>

---

**19** The dataset will include Frame, 1995; Weiershäuser & Novotny, 2019, 2020, 2022. Weiershäuser & Novotny, 2019, 2020, and 2022 will appear in the newly-established Royal Inscriptions of the Neo-Babylonian Empire (RINBE) series, which is co-edited by Radner and Frame, managed by Novotny and published by Eisenbrauns. Some of the inscriptions to appear in those three volumes have already been published in Schaudig, 2001; and Da Riva, 2009, 2012, 2013.

**20** The “dynastic” numbering follows that of the Royal Inscriptions of Mesopotamia, Babylonian Periods (RIMB) publications of the now-defunct Royal Inscriptions of Mesopotamia (RIM) Project directed by A. Kirk Grayson at the University of Toronto. “Babylon 1” = Kassite Period (1595–1155 BCE); “Babylon 2” = Second Dynasty of Isin (1157–1026 BCE); “Babylon 3” = Second Dynasty of the Sealand (1025–1005 BCE); “Babylon 4” = Bazi Dynasty (1004–985 BCE); “Babylon 5” = Elamite Dynasty (984–979 BCE); “Babylon 6” = Uncertain Dynasties (978–626 BCE); “Babylon 7” = Neo-Babylonian Dynasty (625–539 BCE); “Babylon 8” = Akkadian inscriptions of the Persian Period (538–330 BCE); “Babylon 9” = Macedonian rulers of Mesopotamia (currently no inscriptions known); and “Babylon 10” = Seleucid era (305–64 BCE).

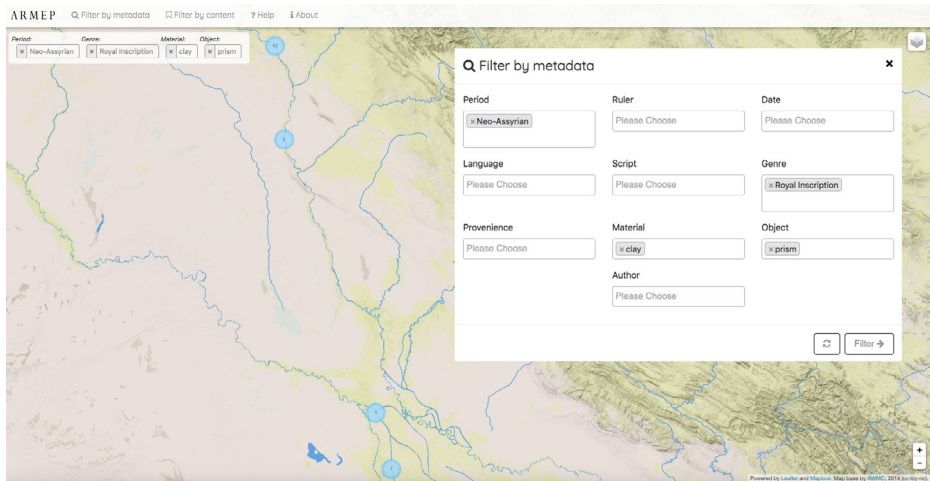
**21** [<http://oracc.museum.upenn.edu/ribo/corpus>].

**22** Frans van Koppen of the Institut für Assyriologie und Hethitologie at LMU Munich is currently overseeing the creation of editions of the inscriptions of the First Dynasty of Babylon, including the famous Law Code of Hammurabi stele. Grant Frame is working on the Kassite material.



genre, language, material support, object type, period, provenience, ruler, and script) and content (translations, transliterations, lemma, and cuneiform signs).<sup>25</sup>

The following is an example of metadata filtering: when a user selects “Neo-Assyrian” for the period, “Royal Inscription” as the genre, “clay” as the material, and “prism” as the object, the map (using the current dataset) displays fifty-five (composite) texts originating from five different sites (Ashur, Babylon, Nimrud, Nineveh, and Sippar (Figure 11.2).



**Figure 11.2:** Example of ARMEP filter by metadata results

Alternatively, users can search the contents of the texts themselves. For example, when a user searches for the Akkadian lemma “anzû” (a mythical lion-headed eagle) and “lābu” (a word for lion), the map (using the current dataset) displays twenty-three (composite) texts found at seven sites (Ashur, Babil, Babylon, Nimrud, Nineveh, Uruk, and Zinçirli) (Figure 11.3).

<sup>25</sup> The current dataset includes: RIAo (866 Assyrian inscriptions from the third millennium BCE to 745 BCE), RIBo (209 Babylonian inscriptions from 1157–64 BCE), RINAP (674 Assyrian inscriptions from 744–612 BCE), SAAo (4888 Neo-Assyrian archival texts published by the Helsinki-based Neo-Assyrian Text corpus project), and Suhu (33 inscriptions from the ninth century BCE).



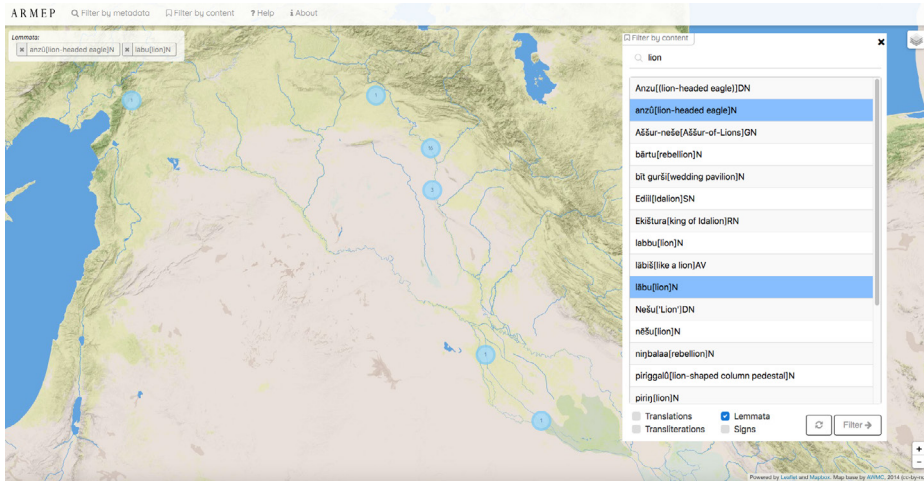


Figure 11.3: Example of ARMEP filter by content results

The lemmatized texts on Oracc can be accessed from the “Item View” pop-up box, which is accessible through the “Cluster Overview” pop-up (Figure 11.4).

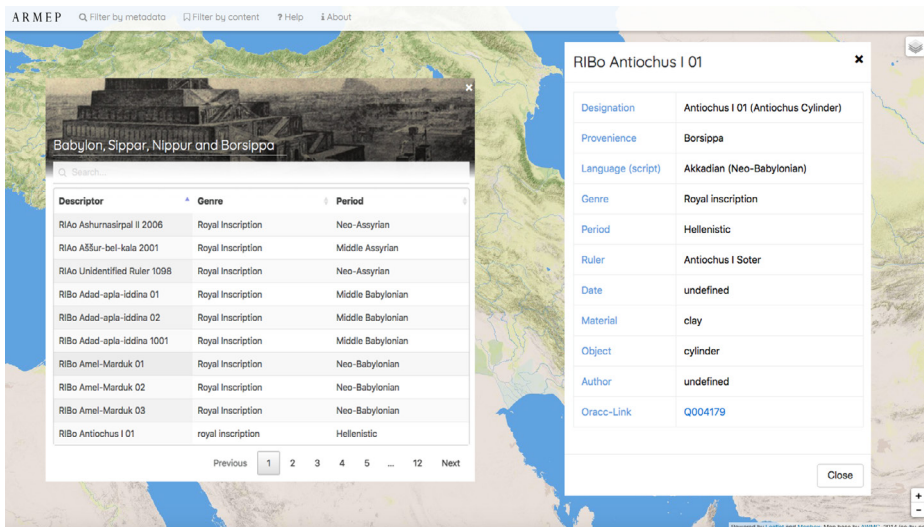


Figure 11.4: Sample “Cluster Overview” (left) and “Item View” (right)

ARMEP's architectural structure was deliberately designed so that it can be easily cloned and adapted to meet the needs of other geo-referenced corpora. At LMU, its versatility has already been demonstrated by its adaptation for two other datasets.<sup>26</sup>

ARMEP does not intend to produce digital maps. This open-access, web-based tool is not envisioned as being just a map of Middle Eastern polities, but rather as an interface that contextualizes ancient sources geographically by their find-spots and that allows users direct access to lemmatized editions of the ancient texts found at those sites. This interface is specifically designed to break away from the boring, traditional "catalogue"-style of corpus organization, which is not particularly informative or interesting to non-specialists.

## 11.4 Methodological Problems and Technical Issues

For the most part, the present authors experienced relatively few methodological and/or technical issues with the development of OIMEA and ARMEP.

In terms of corpus building, most potential problems had been ironed out several years earlier as the software used for the Oracc platform has been in continual use since 2007, when it was developed on the basis of the electronic Pennsylvania Sumerian Dictionary.<sup>27</sup>

However, a few relatively minor problems still exist. For OIMEA and its sub-projects, these range from relatively straightforward technical issues to complex challenges:

1. printing easy-to-read editions from the Oracc web interface is not (yet) user-friendly nor are the results (yet) elegant;
2. multi-language support: the Oracc web interface does not properly handle multi-language translations of texts and does not (yet) support right-to-left scripts such as Aramaic, Arabic or Hebrew;
3. geo-referencing place name data in the glossaries is not yet possible;
4. disambiguation of namesakes: Oracc's lemmatizer is currently not able to reliably disambiguate namesakes of people and places, even when additional information is provided for a name's guide word (that is, a word's primary meaning in the glossary);
5. extensive manual review of auto-lemmatized data is still required as the Oracc processor is unable to fully and accurately assess the citation form (the dictionary form of words), the sense (the meaning of the word in context), and the transcription (how the word was pronounced).

---

<sup>26</sup> Digitizing Ancient Near Eastern Seals and Sealings [<http://www.diganes.gwi.uni-muenchen.de/>] and Dynastie im Raum: Die Grabstätten der Habsburger 991–1996 [<http://www.habsburg.gwi.uni-muenchen.de/>].

<sup>27</sup> ePSD [<http://psd.museum.upenn.edu/epsd1/index.html>].

Most of these issues are technical rather than conceptual, especially the handling of print output (1).<sup>28</sup> Tinney aims to resolve multi-language support (2) by late 2018.<sup>29</sup> The ability to geo-reference place names (3) will be incorporated into the glossary generation as part of the development of ARMEP 2.0 in late 2018. The fix is to simply add a Pleiades<sup>30</sup> ID field to entries in the glossary of names; the longitude and latitude coordinates will then be retrieved automatically from Oracc’s LMU Munich-based Geonames project.<sup>31</sup> The disambiguation of namesakes (4) is needed if one wants to use Oracc-based data to create prosopographies.<sup>32</sup> In order to achieve this, the system processor used for glossary validation needs to be fine-tuned so that the checker looks for 100% matches between lemmatized data in the source files and the corresponding glossary file.<sup>33</sup> The current workaround is to add underscores to the citation form; for example, “Tukulti-apil-Ešarra[Tiglath-pileser III, king of Assyria]RN” becomes “Tukulti-apil-Ešarra\_III[Tiglath-pileser III, king of Assyria]RN”.<sup>34</sup> The last issue, the need for the manual review of auto-lemmatized data (5), is a thorny problem that is likely to remain unresolved for some time as it requires refinement to the Oracc logic processor.<sup>35</sup>

---

**28** Oracc’s “Print text” function is now set up to handle projects with multi-language translations. On screen, the print option displays the editions in a readable format, but when the text is printed on paper (or to PDF) the results are less than desirable, as the multi-column format is not properly handled.

**29** The only outstanding problem is that translation languages are displayed alphabetically by ISO 639-1 language code. Thus, English, which is the primary language of OIMEA projects, may not always be the default translation language in the Oracc pager when other translation languages are used.

**30** Pleiades [<http://pleiades.stoa.org/>]. Various members of the OIMEA team have been volunteer content contributors of Pleiades since 2015 and they have added over 1,100 place resources for Assyrian and Babylonian cities, city gates, city walls, palaces, and temples. Over the course of 2018, we plan to add many more place resources so that ARMEP’s and Pleiades’ geo-referenced data are fully compatible.

**31** [<http://oracc.ub.uni-muenchen.de/geonames/hub.html>].

**32** Heather D. Baker (University of Toronto) is currently working on such a project, the Prosopography of the Neo-Assyrian Empire online (PNAo) [<http://oracc.museum.upenn.edu/pnao/index.html>].

**33** For information about lemmatizing Akkadian and Sumerian texts on Oracc, see [<http://oracc.museum.upenn.edu/doc/help/languages/index.html>]; [<http://oracc.museum.upenn.edu/doc/help/lemmatizing/index.html>]; and [<http://oracc.museum.upenn.edu/doc/help/languages/akkadian/index.html>].

**34** The Oracc processor cannot properly analyze citation forms (CF) with marginally different guide words (GW). For example, Tukulti-apil-Ešarra[Tiglath-pileser I, king of Assyria]RN, Tukulti-apil-Ešarra[Tiglath-pileser II, king of Assyria]RN, and Tukulti-apil-Ešarra[Tiglath-pileser III, king of Assyria]RN are too similar to be disambiguated by the lemmatizer. The tolerance setting needs to be readjusted. For some information on Oracc CFs and GWs, see [<http://oracc.museum.upenn.edu/doc/help/lemmatizing/primer/index.html>].

**35** Oracc’s auto-lemmatizer function is not publically (or privately) documented. The criteria by which lemmatizer selects GWs (or senses) for CFs is not known to the authors. It is usually the GWs/sences, not the CFs (part of speech [POS] or normalization [NORM]), that require manual correction.

As for the ARMEP map interface, there were some issues at the beginning of development because the VerbaAlpina and Oracc data formats were not readily compatible. In addition, the catalogues, glossaries, transliterations, translations, and lists of signs could not be exported from Oracc. The solution<sup>36</sup> was simply to make Oracc catalogue, glossary, transliteration, and translation data available in JavaScript Object Notation (JSON), under a CCO or public domain license,<sup>37</sup> a feature that Tinney implemented in early 2017. For ARMEP 1.0, no further compatibility issues were encountered.

## 11.5 Future Prospects

Over the next five years (2018–2022), OIMEA's content will be expanded to incorporate inscriptions written in scripts other than cuneiform: Aramaic and Luwian are on top of the list. In addition, its lemmatized contents and glossaries will be improved, especially by standardizing the information in glossaries across its numerous sub-projects.

During 2018, ARMEP 2.0 is being developed and will be released at the end of the year (or in 2019). The new version of that open-access web interface will feature a gazetteer mode that will display places (including cities, villages, temples, mountains, and bodies of water) mentioned in ancient sources whose coordinates are known with a reasonable degree of certainty. This will substantially expand the information displayed in ARMEP 1.0, which shows texts according to their find-spots. This gazetteer function will display all geo-referenced places named in the defined text corpus (currently about 6,700 texts). For example, if a user clicks on the “View Places in Text” link of a 7<sup>th</sup> century BCE Assyrian inscription (e.g., the “Final Edition” of the Annals of Sennacherib), the map will show all of the cities that that king claims to have conquered and destroyed, as well as cities from whose rulers tribute was received. This innovative and dynamic geo-referenced rendering, which visualizes data in an easy-to-digest manner never before used in ancient Near Eastern studies, will further enhance the accessibility and usability of geographical information mentioned in cuneiform sources beyond specialist academics to casual or inexperienced users, including beginner students and members of the general public.

---

<sup>36</sup> The Englmeier brothers (with input from Lücke) and Tinney easily resolved the compatibility issue, as well as Oracc's lack of exportability.

<sup>37</sup> For further information, see [<http://oracc.museum.upenn.edu/doc/opendata/index.html>].

## Bibliography

- Da Riva, R. (2009). The Nebuchadnezzar Rock Inscription at Nahr el-Kalb. In A.M. Maïla-Afeiche (Ed.), *Le site de Nahr el-Kalb* (Bulletin d'Archéologie et d'Architecture Libanaises, Hors-Série 5) (pp. 255–301). Beirut: Ministère de la Culture: Direction Générale des Antiquités.
- Da Riva, R. (2012). *The Twin Inscriptions of Nebuchadnezzar at Brisa (Wadi esh-Sharbin, Lebanon): A Historical and Philological Study* (Archiv für Orientforschung, Beiheft 32). Vienna: Institut für Orientalistik der Universität Wien.
- Da Riva, R. (2013). *The Inscriptions of Nabopolassar, Amel-Marduk and Neriglissar* (Studies in Ancient Near Eastern Records 3). Boston: De Gruyter.
- Frame, G. (1995). *Rulers of Babylonia: From the Second Dynasty of Isin to the End of Assyrian Domination (1157–612 BC)* (Royal Inscriptions of Mesopotamia, Babylonian Periods 2). Toronto: University of Toronto Press.
- Frame, G. (2019). *The Royal Inscriptions of Sargon II, King of Assyria (721–705 BC)* (Royal Inscriptions of the Neo-Assyrian Period 2). Winona Lake: Eisenbrauns. Manuscript in preparation.
- Grayson, A.K. (1987). *Assyrian Rulers of the Third and Second Millennia BC (to 1115 BC)* (Royal Inscriptions of Mesopotamia, Assyrian Periods 1). Toronto: University of Toronto Press.
- Grayson, A.K. (1991). *Assyrian Rulers of the Early First Millennium BC I (1114–859 BC)* (Royal Inscriptions of Mesopotamia, Assyrian Periods 2). Toronto: University of Toronto Press.
- Grayson, A.K. (1996). *Assyrian Rulers of the Early First Millennium BC II (858–745 BC)* (Royal Inscriptions of Mesopotamia, Assyrian Periods 3). Toronto: University of Toronto Press.
- Grayson, A.K. & Novotny, J. (2012). *The Royal Inscriptions of Sennacherib, King of Assyria (704–681 BC). Part 1* (Royal Inscriptions of the Neo-Assyrian Period 3/1). Winona Lake: Eisenbrauns.
- Grayson, A.K. & Novotny, J. (2014). *The Royal Inscriptions of Sennacherib, King of Assyria (704–681 BC). Part 2* (Royal Inscriptions of the Neo-Assyrian Period 3/2). Winona Lake: Eisenbrauns.
- Leichty, E. (2011). *The Royal Inscriptions of Esarhaddon, King of Assyria (680–669 BC)* (Royal Inscriptions of the Neo-Assyrian Period 4). Winona Lake: Eisenbrauns.
- Novotny, J. & Jeffers, J. (2018). *The Royal Inscriptions of Ashurbanipal (668–631 BC), Aššur-etel-ilāni (630–627 BC) and Sîn-šarra-iškun (626–612 BC), Kings of Assyria. Part 1* (Royal Inscriptions of the Neo-Assyrian Period 5/1). Winona Lake: Eisenbrauns.
- Novotny, J. & Jeffers, J. (2019). *The Royal Inscriptions of Ashurbanipal (668–631 BC), Aššur-etel-ilāni (630–627 BC) and Sîn-šarra-iškun (626–612 BC), Kings of Assyria. Part 2* (Royal Inscriptions of the Neo-Assyrian Period 5/2). Winona Lake: Eisenbrauns. Manuscript in preparation.
- Salvini, M. (2008–12). *Corpus dei Testi Urartei* (4 vols.). Rome: CNR/Istituto di studi sulle civiltà dell'Egeo e del Vicino Oriente.
- Schaudig, H.P. (2001). *Die Inschriften Nabonids von Babylon und Kyros' des Großen samt den in ihrem Umfeld entstandenen Tendenzschriften. Textausgabe und Grammatik* (Alter Orient und Altes Testament 256). Münster: Ugarit-Verlag.
- Schmitt, R. (2009). *Die altpersischen Inschriften der Achaimeniden: Editio minor mit deutscher Übersetzung*. Wiesbaden: Reichert Verlag.
- Tadmor, H. & Yamada, S. (2011). *The Royal Inscriptions of Tiglath-pileser III (744–727 BC) and Shalmaneser V (726–722 BC), Kings of Assyria* (Royal Inscriptions of the Neo-Assyrian Period 1). Winona Lake: Eisenbrauns.
- Weiershäuser, F. & Novotny, J. (2019). *The Royal Inscriptions of Amel-Marduk (562–560 BC), Neriglissar (560–556 BC), and Nabonidus (555–539 BC), Kings of Babylon* (Royal Inscriptions of the Neo-Babylonian Empire 3). Manuscript in preparation.

- Weiershäuser, F. & Novotny, J. (2020). *The Royal Inscriptions of Nabopolassar (625–605 BC), King of Babylon, and Nebuchadnezzar II (604–562 BC), King of Babylon, Part 1* (Royal Inscriptions of the Neo-Babylonia Empire 1). Winona Lake: Eisenbrauns. Manuscript in preparation.
- Weiershäuser, F. & Novotny, J. (2022). *The Royal Inscriptions of Nebuchadnezzar II (604–562 BC), King of Babylon, Part 2* (Royal Inscriptions of the Neo-Babylonia Empire 2). Winona Lake: Eisenbrauns. Manuscript in preparation.

Sébastien Biston-Moulin and Christophe Thiers

## 12 The Karnak Project: A Comprehensive Edition of the Largest Ancient Egyptian Temple

**Abstract:** This article is concerned with the technical and methodological challenges encountered during a project to comprehensively document the inscriptions of the largest ancient Egyptian temple. This project aims to produce a complete inventory, and editing of, primary textual sources written in several varieties of the ancient Egyptian language and script: hieroglyphs, hieratic and demotic. The issues discussed concern the implementation of the digital tool, the need for a network of collaborators in order to process the large volume of documentation, and the need to identify digital solutions to preserve textual data from the Egyptian site. Finally, the lexicographic aspect of the project is discussed.

**Keywords:** ancient Egypt, Karnak temples, digital hieroglyphic corpora, high resolution orthophotographs, heritage preservation

### 12.1 Introduction

For nearly two millennia the temples of Karnak were one of the religious and political capitals of ancient Egypt. Today, they form an archaeological area of 25 hectares, where thousands of inscriptions, scenes and inscribed objects are preserved or have been discovered on-site. The temple consists of a main complex with a double east-west and south-north axis (Figure 12.1) dedicated to the divinity Amun-Re who, among other prerogatives, guaranteed the rightful transmission of royalty. This complex has therefore received special attention from those who attempted to, or actually gained, power, each ruler seeking to leave his contribution in the temple of the “father” from whom he derived part of his legitimacy to govern. In addition, various temples dedicated to other deities, such as Ptah, Khonsu, and Osiris, are included in the main temple’s enclosure. The hieroglyphic inscriptions of this complex range from 2000 BCE to the first century CE.

Despite the obvious historical, religious and linguistic importance of these documents, the publication of the lexical and iconographic data of this vast sanctuary was, until recently, far from complete. No compilation, index or glossary had been produced to extract the content of these documents.

---

Sébastien Biston-Moulin, CNRS, UMR 5140, Archéologie des sociétés méditerranéennes

Christophe Thiers, CNRS, USR 3172, Centre Franco-Égyptien d’Étude des Temples de Karnak

 © 2018 Sébastien Biston-Moulin and Christophe Thiers

This work is licensed under the Creative Commons Attribution 4.0 International license (CC BY 4.0)



**Figure 12.1:** Main axis of the temple of Amun-Ra at Karnak (© CNRS-CFEETK)

Thanks to the presence of a permanent CNRS team on site,<sup>1</sup> the objective of the *Karnak* project, initiated in 2013, was primarily to collect this unique amount of epigraphic material. This objective immediately raised the question of the organization of such documentation. How to collect, in an optimal way, these inscriptions that span millennia and use different writing systems, and subsequently, how to disseminate the richness of their contents as widely as possible? Since many of the inscriptions are still known only through hardcopies that are sometimes very old (mid-19<sup>th</sup> century), or remain unpublished, how to provide a level of documentation to be used both for the edition of primary sources and for research? Faced with a constantly deteriorating heritage, despite all the attention given to these monuments, what is the best way to sufficiently document, and thereby preserve, the information as it stands today, in the event of a deterioration of these reliefs?

To address these difficulties, we have chosen to build a comprehensive corpus of the primary sources from the site that would collect all the published and unpublished information concerning these inscriptions, as well as high-resolution photographs

<sup>1</sup> French-Egyptian Centre for the Study of the Temples of Karnak (CFEETK – CNRS, USR 3172).



serving autoptic reading. It was also necessary for this corpus to be georeferenced to set each inscription within its textual and iconographic context in the temple. Finally, in order to provide access to the content of the inscriptions, it was necessary to implement solutions for lexical analysis.

This corpus has obviously taken the form of a digital tool used both for editing hieroglyphic texts and for disseminating them.

## 12.2 Towards an Interactive Corpus of Primary Sources in Ancient Egyptian

### 12.2.1 Fieldwork and Implementation of the Tools

The first problem we faced was having a tool that could support the documentation related to a language (ancient Egyptian) that uses a figurative writing (hieroglyphs) with a set of signs having a potentially infinite number of graphic variants, without punctuation. The challenge was even wider, since a significant part of the Karnak temples' inscriptions (more than 10,000 in total) use further writing systems such as hieratic and demotic.

Since the 1990s, projects encoding hieroglyphic texts from a particular corpus or a language stage have multiplied, with the aim of producing lexicographic or morphological analysis tools.<sup>2</sup> However, none of these tools was available for reuse and none seemed to be suitable for producing a reference edition of these texts. The sheer number of documents that needed to be processed was also an obstacle to overcome. It was therefore necessary to develop an *ad hoc* tool meeting the specific objectives of the project.

Hosted by the Huma-Num service grid, which aims to facilitate the digital turn in humanities and social sciences in France, this tool allows the project team to compile the corpus of the inscriptions of Karnak (Figure 12.2).<sup>3</sup>

Once the application had been implemented, the next problem encountered was the lack of an exhaustive inventory of the epigraphic documentation of the

---

<sup>2</sup> To mention only the main ones: *Thesaurus Linguae Aegyptiae* (Hafemann & Dils, 2013) [<http://aaw.bbaw.de/tla/>] and Online Ramses (Polis, Honnay, & Winand, 2013; Polis & Winand, 2013) [<http://ramses.ulg.ac.be/>].

<sup>3</sup> [<http://sith.huma-num.fr/karnak>].

Karnak complex. The project team<sup>4</sup> proceeded to document the complex monument by monument, wall by wall, object by object. The process entails the cataloguing of scenes, objects and inscriptions in a common reference system and the creation of bibliographic records when they have already been published or mentioned. Every text receives a unique identification number (KIU: Karnak Identifiant Unique) that works as a reference throughout the project and enables the creation of URIs (Uniform Resource Identifier) for the inscriptions.

The screenshot shows the SITH Karnak project interface. At the top, there is a navigation bar with the project name and a search bar. Below the navigation bar, the main heading reads "Apparition de la statue royale Scène 4 - (KIU 1098)". The breadcrumb trail indicates the location: "Karnak / Chapelle blanche / Piliers / Pilier 1.n".

The main content area displays a photograph of a stone relief (Cftek 71871) on the left. To the right, the following information is provided:

- DATATION:** XII<sup>e</sup> dynasty / Senusret I
- MATERIAL:** Calcaire.
- Inscription**
- La statue du roi**
  - 1 hr ḥb-maut
  - 2 naut ḥfy Ḥpr-ky-R'
- Amon-Ré**
  - 3 Jmn-R' d'ef ḥb
- Sous Amon-Ré**
  - 4 nry (Jmn) dw ḥb d'f wꜣꜣꜣ nb ꜣꜣ Jbꜣf ḥ' hr Wꜣt-Ḥr-ꜣꜣꜣꜣ dw ḥb

Below the main image, there are two smaller images with their respective IDs: Cftek 71871 (2004) and Cftek 58208 (2001). The interface also includes a "type" dropdown menu and a "date" dropdown menu.

Figure 12.2: A scene and its inscriptions from the White Chapel of Senusret I (ca. 2000 BCE)<sup>5</sup>

The second issue was the integration of different language stages (Middle Egyptian, Late Egyptian, Ptolemaic) and writing systems (hieroglyphs, hieratic, demotic). Two teams, one from the University of Oxford (Dr. Elizabeth Frood and Chiara

<sup>4</sup> Thanks to substantial funding in the form of a “Laboratoire d’Excellence” called Archimede, for a seven-year period (2013–2019), it has been possible to bring together a team that has grown over the years from five to seven people. Since 2013, 37 authors contributed to the project: Dr. Ali Abdelhalim Ali, Romane Betzeze, Silke Cassor-Pfeiffer, Dr. Léo Cagnard, Dr. Marion Claude, Dr. Laurent Coulon, Edwin Dalino, Dr. Gabriella Dembitz, Dr. Didier Devauchelle, Dr. Abraham Fernandez Pichel, Tiphaine Fignon, Elsa Fournie, Dr. Marc Gabolde, Dr. Luc Gabolde, Dr. Mohamed Gamal Rashed, Maeva Gervason, Mounir Habachy, Fanny Hamonic, Dr. Jérémy Hourdin, Marie-Paule Jung, Dr. Charlie Labarta, Dr. Françoise Labrique, Dr. Cédric Larcher, Mélie Louys, Dr. Dina Metawi, Dr. Elena Panaite, Anne-Hélène Perrot, Dr. Renaud Pietri, Dr. René Preys, Dr. Émeline Pulicani, Dr. Mohamed Raafat Abbas, Dr. Laurie Rouviere, Chiara Salvador, Dr. Anaïs Tillier, Dr. Ghislaine Widmer and the present authors.

<sup>5</sup> [http://sith.huma-num.fr/karnak/1098].

Salvador) and the other from the University of Lille (Dr. Didier Devauchelle and Dr. Ghislaine Widmer), thus joined the project for the integration of hieratic and demotic documentation. While hieroglyphic texts are entered using a hieroglyphic word processor (Rosmorduc, 2014) and a font adapted to Karnak's inscriptions, we have chosen, in accordance with these two partners, to use facsimiles embedded in the interface as a medium for the hieratic and demotic texts.

All inventoried documents have been then organized topographically, and the decorations of the monuments have been arranged hierarchically by section, wall, register, and so on. This work enables immediate contextualization of the different texts in the temples, and the ability to move easily from one to those around it.

To avoid restricting the work carried out on the project's online interface, a first volume of the inventory of monuments, objects, scenes and inscriptions of the temples of Karnak, gathering all information collected in the framework of the project, was published in 2016 (Biston-Moulin, 2016). This inventory will be periodically updated in the coming years.

The project seeks to be as thorough as possible and includes data from the *Cachette of Karnak*, a database which has been developed by the French Institute for Oriental Archaeology (IFAO) and the CNRS since 2006 (Coulon & Jambon, 2016) devoted to about a thousand statues and objects unearthed at the same location in the Karnak temple at the beginning of the 19<sup>th</sup> century, and which are now kept in various museums around the world.<sup>6</sup>

### 12.2.2 Production and Dissemination of Reference Documents

An additional technical difficulty was managing a large amount of photographic data. In addition to the text edition, one of the objectives of the *Karnak* project is to produce a complete photographic record of the inscriptions of the temples. High-resolution photographs had to accompany the publication of the inscriptions in the project interface. We have chosen to transfer the management of these files to Nakala, a service also provided by the Huma-Num research infrastructure for raw data,<sup>7</sup> which grants unlimited storage for this photographic coverage. These files are thus separated from the publication interface of hieroglyphic texts, but can be accessed at any time as a reference document. Facsimiles of inscriptions and other archival documents may also be made available in this way.

---

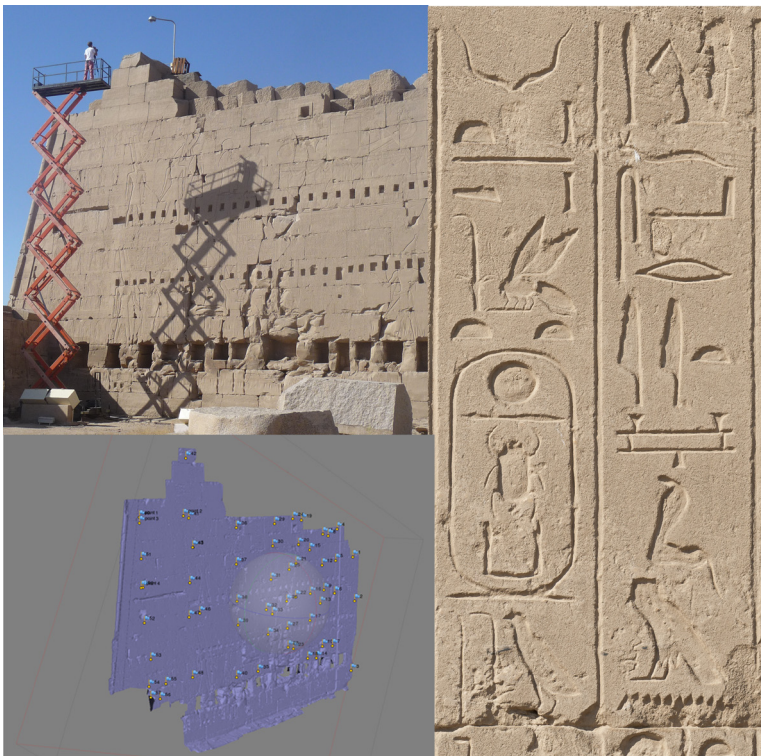
<sup>6</sup> The database of the *Cachette* is available at [<http://www.ifao.egnet.net/bases/cachette/>]; for the implementation of its data in the *Karnak* project, see [<http://sith.huma-num.fr/karnak/3312>].

<sup>7</sup> [<https://www.nakala.fr/>].

These high-resolution photographs are given metadata and distributed online. The metadata associated with these documents are interoperable (RDF/Sparql) to allow searches and ensure both data accessibility and reliability over time.

The user is therefore consistently provided with tools for source criticism. The photographs also allow access to the palaeography of texts and to the relationship between these and the decors. The whole collection of archival photographs of the CFEETK, which cover nearly 150 years of work in the temple (1870–2018), are also associated with the relative documents through the Nakala repository, showing whether a text or a decoration is in a different state of preservation or context than in the past. Approximately 30,000 photographs illustrating the inscriptions are available at this stage.

In order to obtain this coverage, a photographic campaign was established. Photogrammetric techniques are consistently used to produce reliable high-resolution orthophotographs of temple walls and objects in a limited time (Figure 12.3; Tournadre et al., 2017).



**Figure 12.3:** Orthophotographic survey, data processing in Photoscan and detail of orthophotography of an inscription several meters high acquired by means of this technique

This exhaustive photographic coverage, the first made for Karnak, is also intended to preserve the textual and iconographic heritage of the temple as it stands today. Climate and anthropogenic degradations are to be feared and the disappearance of a relief, or part of it, is an irreplaceable loss. This may be counteracted with exploitable, high-resolution photographs, making this programme an absolutely crucial step towards the heritage preservation of the largest temple of Egypt, and should be encouraged on a broader level for all Egyptian sites. Although this method is very fast to implement and the work is progressing rapidly, one of our concerns is that it may be difficult to complete the photographic coverage within the timeframe of the project funding.

### 12.2.3 From Plain Text to Indexed Interactive Text

The last technical issue we will discuss here is linked to the encoding of texts in ancient Egyptian. An interactive text in which the user can search, browse and see the contents with indexes has always been one of the central ideas of the project. Because of the complexity of the corpus itself, and the priority given to the acquisition and publication of primary sources on site, this step of the project could not be undertaken before 2015. In order to achieve this objective for hieroglyphic inscriptions, it was necessary to develop an indexation system flexible enough to process a very large quantity of lexical data, but also sufficiently detailed to allow a careful lexical analysis of inscriptions.

Because of the partial knowledge of the ancient Egyptian vocabulary, we obviously needed a partner at this stage to undertake this lexical exploitation of Karnak's data. We turned to the dictionary project of the University of Montpellier *VÉGA – Vocabulaire de l'Égyptien Ancien* led by Fr. Servajean, which aims to produce the first updated dictionary of ancient Egyptian in French since Jean-François Champollion.<sup>8</sup> The richness of the data collected by the *Karnak* project was greatly valuable for the production of a dictionary, thus facilitating the partnership between the two projects.

In 2015 we were therefore able to develop a new tool called “Système d'Indexation des Textes Hiéroglyphiques”, for indexing hieroglyphic, hieratic and demotic texts. This programme is designed to create lists of words, theonyms, toponyms, ethnic names and cult places, anthroponyms and names of kings from the contents of the corpus. It then detects possible attestations and allows the creation of indexes, classified both chronologically and topographically in the temple. To date, thanks to the indexing work, several hundred thousand attestations of identified terms and contexts are proposed. This application is used to transform the plain text entered by the members of the project into an interactive text indexed by the detection of

---

<sup>8</sup> [<http://vega-vocabulaire-egyptien-ancien.fr/>].

the occurrences and morphological features of the elements of the sentence. Each possible attestation is then manually validated or rejected. The result is an annotated corpus that allows very detailed searches or compilations based on chronology, grammatical features or context of use (Figure 12.4).

The screenshot shows the SITH website interface. At the top, there is a search bar with the text 'Rechercher' and a dropdown menu. Below the search bar, there are navigation tabs: 'Titulature', 'Toponymes', 'Thésonymes', and 'Anthroponymes'. The main content area is titled 'Vocables' and displays the search results for the word 'nsyt'. The word is shown in its hieroglyphic form and is followed by the text « Royauté ». Below this, there is a section for 'Vocabulaire' with a grid of characters and a list of attestations. The first attestation is for 'nsyt' (substantif) with 279 attestations. Below this, there are three specific attestations listed: KIU 1034, KIU 1038, and KIU 1061, each with its corresponding hieroglyphic transcription and a small image of the inscription fragment.

Figure 12.4: The world *nsyt* « Kingship » in the inscriptions of Karnak<sup>9</sup>

To broaden the dissemination of this compiled data, and reach a different audience from the online interface, a first volume of the Glossary of the Inscriptions of Karnak dedicated to the vocabulary was published in 2017 (Biston-Moulin, 2017). It includes about 100,000 word attestations spread over a little more than 2,000 years of use in Karnak. In the coming years, it is intended to periodically update this volume, giving access to an ever-increasing number of texts, and identified terms, attestations and contexts.

Much remains to be done in order to complete and enrich the indexation of the inscriptions collected as part of the constitution of the corpus of Karnak texts. One of the objectives will be to make the whole corpus fully interoperable (TEI/EpiDoc) in order to increase its dissemination and allow the total or partial reuse of Karnak texts and indexed lexical data.

One of the main difficulties in advancing this part of the project is the absence of a recent reference work or compilation mainly for lexicon, anthroponyms or toponyms. While the production of a lexicon of ancient Egyptian will hopefully be achieved

<sup>9</sup> [http://sith.huma-num.fr/vocabulaire/111].

as a result of the progress of projects dealing with dictionaries, the production of an updated geographical gazetteer of toponyms attested in Egyptian inscriptions remains a remarkable desideratum.<sup>10</sup>

### 12.3 Progress and Prospects

These are a few aspects of the main technical and methodological challenges that the *Karnak* project has had to overcome in the course of the production, still in progress, of the largest corpus of hieroglyphic texts freely available online.

Through the choices made at the outset of the project, then during its development, and the technical solutions developed along the way, five years after its launch the *Karnak* project has collected, organized and edited more than 10,000 hieroglyphic, hieratic and demotic inscriptions. Its online interface available in French, English and Arabic has received more than 4,000,000 visitors.

The edition of the Karnak project corpus will be completed in the coming years and our attention is now turning to the future of the data collected in the course of this digital epigraphy project. All the photographs are already stored and distributed via a system ensuring their long-term preservation. All of the textual data will be released in Open Access under a Creative Commons license.<sup>11</sup>

Beyond the difficulty in finding reference tools for ancient Egyptian, one of the unresolved questions of the project is the catalogue of the graph variants of the hieroglyphic signs composing the various attestations of one term. This would be an extremely valuable addition to the existing data, but will probably require the implementation of specific tools that have yet to be defined for the project. This dimension obviously involves the photographic documentation that we have already collected, but also the work on the facsimiles. Even though this activity has been carried out since the beginning of the project, its progress is very slow, because of the time needed to produce such documents.

The technical solutions and methodological choices adopted in the development of a digital epigraphy project on the largest Egyptian temple could naturally function as a foundation for the extension of the project beyond the Karnak temples. Integrating texts from other Egyptian sites or thematic corpora would certainly be the right step

---

<sup>10</sup> These geographical names obviously concern territories, cities, temples, and monuments all over Egypt, but they also include numerous Asian and African territories, localities and ethnics whose names have been recorded in Egyptian texts. A few references to European place names such as the name of the city of Rome (*Hrm*) engraved in hieroglyphic inscriptions of Emperor Augustus in the temple of Opet at Karnak may also be found: [<http://sith.huma-num.fr/toponyme/33>], [<https://www.nakala.fr/nakala/data/11280/e24901f5>].

<sup>11</sup> Attribution-ShareAlike 4.0 International (CC BY-SA 4.0) [<https://creativecommons.org/licenses/by-sa/4.0/>].

to open the way for a much larger collection of inscriptions in ancient Egyptian, overcoming the obstacles discussed here and benefiting from the flexibility and advantages of digital epigraphy for the edition, analysis and publication of sources in ancient Egyptian.

## Bibliography

- Biston-Moulin, S. (2016). *Inventaire des monuments, objets, scènes et inscriptions des temples de Karnak*. Montpellier. Retrieved from [<https://halshs.archives-ouvertes.fr/halshs-01329927/document>], 2017/11/1.
- Biston-Moulin, S. (2017). *Glossaire des inscriptions de Karnak I. Le vocabulaire*. Montpellier. Retrieved from [<https://halshs.archives-ouvertes.fr/hal-01549230/document>], 2017/11/1.
- Coulon, L. & Jambon, E. (2016). L'exploitation scientifique de la Cachette de Karnak, de Georges Legrain à nos jours. Essai d'historiographie. In L. Coulon (Ed.), *La Cachette de Karnak. Nouvelles perspectives sur les découvertes de Georges Legrain* (pp. 89–129). Cairo: Ifao.
- Hafemann, I. & Dils, P. (2013). Der Thesaurus Linguae Aegyptiae – Konzepte und Perspektiven. In I. Hafemann (Ed.), *Perspektiven einer corpusbasierten historischen Linguistik und Philologie. Internationale Tagung des Akademienvorhabens „Altägyptisches Wörterbuch“ an der Berlin-Brandenburgischen Akademie der Wissenschaften, 12.–13. Dezember 2011* (pp. 127–143). Berlin: BBAW.
- Polis, S., Honnay, A.-C., & Winand, J. (2013). Building an Annotated Corpus of Late Egyptian. The Ramses Project: Review and Perspectives. In S. Polis & J. Winand (Eds.), *Texts, Languages & Information Technology in Egyptology* (pp. 25–44). Liège: Presses universitaires.
- Polis, S. & Winand, J. (2013). The Ramses project. Methodology and practices in the annotation of Late Egyptian Texts. In I. Hafemann (Ed.), *Perspektiven einer corpusbasierten historischen Linguistik und Philologie. Internationale Tagung des Akademienvorhabens „Altägyptisches Wörterbuch“ an der Berlin-Brandenburgischen Akademie der Wissenschaften, 12. - 13. Dezember 2011* (pp. 81–108). Berlin: BBAW.
- Rosmorduc, S. (2014). *JSesh Documentation*. Retrieved from [<http://jseshdoc.qenherkhopeshef.org>], 2017/11/17.
- Tournadre, V., Labarta, Ch., Megard, P., Garric, A., Saubestre, E., & Durand, B. (2017). Computer Vision in the Temples of Karnak: Past, Present & Future. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 42-5/W1*, 357–364. doi: 10.5194/isprs-archives-XLII-5-W1-357-20



---

## Part II: Providing Access: Portals, Interoperability and Aggregators



Gerfrid G.W. Müller and Daniel Schwemer

## 13 Hethitologie-Portal Mainz (HPM). A Digital Infrastructure for Hittitology and Related Fields in Ancient Near Eastern Studies

**Abstract:** The Hethitologie-Portal Mainz is a not-for-profit, open-access digital infrastructure for Hittitology and related fields of research in Ancient Near Eastern studies. HPM, which was first established in 2001, gives access to an array of interconnected research documents, including critical editions of Hittite cuneiform texts, catalogues, bibliographies, onomastic databases as well as media archives with digital photos, drawings, and 3D models. The HPM community has been constantly growing over the past years and currently comprises more than fifty creators of contents and approximately 3,000 individual human users. User statistics and feedback by peers show that HPM has become essential to Hittitological research. Its digital strategy favours open-source, widespread software and standardized, well-documented data formats in order to ensure long-term sustainability. The absence of low-level, permanent funding opportunities for digital infrastructures in the Humanities in Germany is one of the challenges faced by HPM.

**Keywords:** cuneiform, Hittite, edition, bibliography, digital humanities

### 13.1 Remit and Unique Proposition

The *Hethitologie-Portal Mainz*<sup>1</sup> (Figure 13.1) is one of the main digital infrastructures in Ancient Near Eastern studies. HPM's specific remit is the study of the cultures of ancient Anatolia (Turkey), in particular those of the Late Bronze Age (second half of the second millennium BCE). The kingdom of the Hittites (17<sup>th</sup>–13<sup>th</sup> cent. BCE) plays a prominent role in this period of ancient Near Eastern history. Thus the epigraphic finds from Hittite sites, not least the continuously growing body of cuneiform tablets and fragments (current count approximately 33,000, including unpublished texts) are at the centre of HPM (cf. generally Schwemer, 2017).

---

1 HPM [<http://hethiter.net>].



Figure 13.1: The frontpage of HPM

The move of cuneiform studies as a whole into the online digital age began in the late 1990s in a number of different initiatives, most of which focused on one sub-corpus of cuneiform texts defined by parameters such as provenance, date, language, or text genre; usually these restricted bodies of cuneiform texts also correspond to sub-disciplines of Ancient Near Eastern studies. Hence HPM, the digital infrastructure of Hittitology, forms part of a varied and international landscape of online corpora of cuneiform texts.<sup>2</sup>

Within the range of digital cuneiform online databases, HPM is unique not only with regard to its focus on Hittitology. It also stands out due to its combination of complex philological text editions with related databases including catalogues of epigraphic objects, bibliographies, onomastic indexes, and gazetteers as well as media databases of drawings, photos, and 3D models.

<sup>2</sup> Most importantly: *Ebla Digital Archives* (EbDA: Eblaite, Early Dynastic III period; [http://ebda.cnr.it/]); *Database of Neo-Sumerian Texts* (BDTNS: Sumerian, archival texts, Ur III-period; [http://sefarad.filol.csic.es]); *The Electronic Text Corpus of Sumerian Literature* (ETCSL: Sumerian, literary texts; [http://etcsl.orinst.ox.ac.uk]); *Archives babyloniennes XXe–XVIII siècles av. J.-C.* (ARCHIBAB: Akkadian, archival texts, Old Babylonian period; [http://www.archibab.fr]); *The Neo-Babylonian Cuneiform Corpus* (Nabucco: Akkadian, archival texts, first-millennium Babylonia; [http://nabucco.arts.kuleuven.be]); *Sources of Early Akkadian Literature* (SEAL: Akkadian, literary texts, 3<sup>rd</sup> and 2<sup>nd</sup> millennium BCE; [http://www.seal.uni-leipzig.de]). In contrast to these sites, the *Open Richly Annotated Cuneiform Corpus* (ORACC: [http://oracc.museum.upenn.edu]) is not restricted by language, provenance or text genre, but hosts a range of independent ‘projects’ with various editions of Sumerian and Akkadian texts. A complementary tool to these period- or genre-specific corpora is the database of the *Cuneiform Digital Library Initiative* (CDLI: [https://cdli.ucla.edu]), which aims to provide a complete catalogue of cuneiform tablets and fragments. The CDLI site offers a wide variety of digital tools and materials for Assyriology and has become an essential platform for the publication of photos of cuneiform texts by museums around the world.

## 13.2 Objectives: Innovation, Collaboration, Acceleration

For all its dynamic development, the basic objectives of HPM have not changed since its first inception in 2001. They flow from HPM's design as a digital infrastructure serving the field of Hittitology as a sub-discipline of cuneiform studies.

HPM provides sustainable online access to primary sources in the form of critical editions of texts, transliterations of individual cuneiform manuscripts, and representations of the archaeological objects on which these texts are inscribed, most commonly (fragments of) clay tablets; these representations include digital images (technical drawings and photos) as well as 3D models. HPM's goal is to present the sources in a form that is compliant with the academic standards of Hittitology. The projects associated with HPM also take on an active role in the further development of these academic standards.

In addition to the presentation of sources, HPM provides sustainable online access to research documents of various types, especially catalogue databases, bibliographies, and onomastic indexes. These tools and materials include legacy data collections whose accessibility is preserved by hosting them on HPM.

HPM strives to develop dynamic, digital interconnections between the primary sources, research documents and data collections that form part of the infrastructure. From a current user's perspective, the transition between various components hosted on HPM is already fluid. Frequently, Hittitologists are able to move seamlessly between editions, catalogues, bibliographies, and media databases.

Today, a considerable number of Hittitologists, including junior and postdoctoral researchers, present their data collections on HPM. Thus they reduce the burden of routine tasks for the individual researcher and help to avoid duplicating efforts. The amount of collective research time saved by the digital publication of S. Kořak's *Konkordanz der hethitischen Keilschrifttafeln*<sup>3</sup> and D. Groddek's *Groddeks Liste*<sup>4</sup> is immeasurable; these are only two game changers that have fundamentally transformed Hittitology's working methods. By functioning in this way as a digital platform for individual scholars and research projects of any size, HPM aims to foster collaboration and accelerate research procedures in Hittitology.

The digital medium, which enables authors and creators to update their research data, is especially suitable for any kind of catalogue or data collection requiring growth and modification as knowledge advances and the available sources constantly increase (a typical feature of cuneiform studies). This potential for openness of the digital medium, in contrast to print, also offers an appropriate framework for encouraging the publication of less definitive research efforts and collections of

---

3 [<http://hethiter.net/hetkonk>]

4 [<http://hethiter.net/grodlist>]

raw data. As a matter of principle, the competence to make editorial changes in any document on HPM stays with its creator(s) if they do not decide otherwise.

Last but not least, HPM has also become a space for developing innovative digital research methods and strategies in Hittitology; e.g., the metrological analysis of digitized cuneiform tablets (3D models) as a revolutionary method of palaeography and script classification<sup>5</sup>, or the development of a fully automated digital annotation of transliterated Hittite texts with lexical and morphological metadata<sup>6</sup>.

### 13.3 History and Status Quo 2017

HPM was conceived and created by G. Wilhelm in cooperation with G.G.W. Müller from 2001 onwards in a collaboration between the Academy of Sciences and Literature, Mainz (Academy Programme project *Hethitische Forschungen*, 1961–2015; Wilhelm, 2008; 2015), and Ancient Near Eastern studies at Würzburg University. Initial funding was provided by the Deutsche Forschungsgemeinschaft in 2001–2007 within the framework of the project *Informationsinfrastruktur für digitale Publikation keilschriftlicher Staatsverträge der Hethiter und für darauf bezogene netzbasierte Forschungskooperation* (Würzburg University; Wilhelm, 2013; Müller & Wilhelm, 2015).

In the years 2008–2015, HPM was continuously expanded within the framework of the project *Hethitische Forschungen*, directed by Wilhelm at the Mainz Academy. In that period of time, the most significant extension of the text editions presented on HPM was realized by the following research projects, all funded by the Deutsche Forschungsgemeinschaft: *Digitale Publikation hethitischer Texte: Die Beschwöungsrituale der Hethiter (CTH 390–500)* (Mainz University, 2010–2017; director: D. Prechel); *Hethitische mythologische Texte* (Marburg University, 2005–2008; director: E. Rieken); *Sprachlich-philologische Bearbeitung und digitale Edition der Hymnen und Gebete in hethitischer Sprache (CTH 371–389)* (Marburg University, 2011–2014; director: E. Rieken).

Since 2016, HPM has been an essential component of the digital strategy and publication plan of the Academy Programme project *Das Corpus der hethitischen Festrитуale: staatliche Verwaltung des Kultwesens im spätbronzezeitlichen Anatolien*<sup>7</sup> at the Mainz Academy (2016–2036; directors: E. Rieken and D. Schwemer). HPM as such, however, has no permanent funding arrangement.

Internal user statistics show that the various components of HPM have become an essential everyday tool of Hittitological research. HPM has approximately 3,000

<sup>5</sup> [<http://www.cuneiform.de>].

<sup>6</sup> A prototype of this tool is currently tested within the project HFR – *Das Corpus der hethitischen Festrитуale*, and will be fully operational in 2019.

<sup>7</sup> [<http://www.adwmainz.de/projekte/corpus-der-hethitischen-festrитуale>].

individual human users per annum. About half of HPM's users are based in Germany, ca. 1,000 users are based outside Germany but still within Europe and ca. 500 are outside Europe; a distribution corresponding to the locations of Hittitology at universities worldwide. The sites of HPM have approximately 5.5 million accesses by individual human users per annum. This high number confirms that today “studying Hittite is unthinkable without the Portal” (de Roos, 2007, p. 187).

### 13.4 Organization: A Network of Researchers and Projects

HPM has been recognized by the Deutsche Forschungsgemeinschaft as a research infrastructure<sup>8</sup>. It is a not-for-profit, open-access host and platform. All its content is openly accessible upon publication (“gold open access”). HPM also serves as a gateway to other websites offering research content in Hittitology and related fields; however, the maintenance of (ever changing) external links poses challenges.

HPM is led by a steering committee that, in addition to the two present authors, currently comprises three further specialists in Hittitology (Prechel; Rieken; Wilhelm). The steering committee is tasked with the strategic and technical development of the HPM site. Most importantly, the members of the steering committee liaise with colleagues who create and present content on the HPM site. The committee is assisted in its work by an international scientific advisory board whose members are senior academics and leaders of Hittitological research in their country.

HPM is a platform that offers research projects and individual researchers a sustainable space for presenting and interconnecting digital content such as text editions, media, and data collections. At present, more than fifty colleagues are creating content on HPM either as individuals or as project researchers, with contributions ranging from text editions to bibliographies and geographical databases. The long-term Academy-programme projects *Hethitische Forschungen* (up to 2015) and *Corpus der hethitischen Festrivale* (2016–2036) use HPM as a digital publication platform. Due to their extensive funding periods and the role of their researchers on HPM's steering committee, these two projects have been essential for the maintenance and further development of HPM; this situation will not change for the foreseeable future.

As a platform and portal, HPM considers it very important that the creators of contents and their sources of funding are clearly identifiable and visible on the website, and that clear information is provided on how to refer to and quote from the research documents it hosts. It has been discussed for some years in the HPM community to what extent and in which form older, outdated versions of research documents should be archived and kept available to all users. Indications are that HPM will now move to a public archive solution that provides access to previous

---

<sup>8</sup> See [[http://risources.dfg.de/detail/RI\\_00500\\_de.html](http://risources.dfg.de/detail/RI_00500_de.html)].

versions of some components (e.g., text editions and the *Konkordanz*) if significant changes have occurred. The decision on how their content is presented will, however, always stay with the relevant creators.

Institutionally, HPM is located at the Mainz Academy of Sciences and Literature (Department Hethitologie-Archiv) and also has office space at the Ancient Near Eastern studies section of the Department for Ancient Cultures at Würzburg University. For data storage and retrieval, it uses local servers at the Academy as well as servers of the computing centres of Würzburg University and Mainz University.

## 13.5 Digital Components and Concepts

### 13.5.1 Components of HPM

The heart of HPM is the *Konkordanz der hethitischen Keilschrifttexte* created by Košak. It lists all known Hittite fragments and tablets with their place of discovery, date, joins, and bibliographical references. In addition to its own search interface, the *Konkordanz* can be accessed content-wise through CTH, the *Catalogue of the Texts of the Hittites* (the digital continuation of Laroche, 1971 and supplements). From the *Konkordanz* one can also reach the *Joinskizzen*, which show the placement of the fragments within a fragmented clay tablet, as well as the text editions on HPM.

The reconstruction and online publication of the Hittite texts is one of the main objectives of HPM. In addition to Hittite texts (see section 3 for the most extensive text groups currently available), HPM also includes Old Assyrian texts from Anatolia by K. Hecker and, in the future, documents from the Hurrian cultural area (Nuzi).

For the exploration of research literature, there are several bibliographies that have a different scope than the manuscript-based *Konkordanz* and the passage-based *Groddeks Liste*. The comprehensive *Hethitische Bibliographie* was started by J. Součková (Prague), Müller and Wilhelm for older literature, and has now been maintained for many years by M. Marazzi (Naples) in cooperation with several other colleagues. It is complemented by the *Systematische Bibliographie* (mainly supervised by Součková) and a bibliography of Hittite lexemes (led by Marazzi and N. Bolatti Guzzo in collaboration with other colleagues).

Various onomastic databases for personal names, place names and divine names are, or will, soon be available and can be used for indexing the text editions and other data sets on HPM.

The media archive contains approximately 70,000 photos of Hittite texts, as well as a photo collection of Alalakh texts and photos of Old Assyrian texts. Since 2017, HPM gives access to more than 2,000 3D scans of cuneiform tablets. Viewing of the 3D models is enabled by the programme *Cuneiform WebGLViewer* (Figure 13.2),



which was created by D. Fisseler for HPM, and allows an exact examination and measurement of the surface.<sup>9</sup>



**Figure 13.2:** The WebGLViewer of HPM allows the collation of cuneiform tablets in the web browser and provides several tools for measuring and enhancement

Finally, HPM offers various services, including downloadable fonts and e-books (Studien zu den Boğazköy-Texten; HPM – Materialien) as well as some general information on Hittite history and culture.

### 13.5.2 Open Standards and Widespread Open-Source Software

When, in 2000, Wilhelm and Müller first discussed the development of an information infrastructure for the digital publication of Hittite texts, it was clear from the start that only international standards with the widest possible dissemination could be used for an enterprise of this type.

From the beginning, preference was given to open-source software. As with standards, wide dissemination and large user numbers were important criteria for the choice of software solutions in order to safeguard stability, sustained compatibility and continuous further development. Too much work and effort had been spent on the transitions from Apple II to Atari to DOS and Windows, from 7-bit to 8-bit data format, from the inadequate ASCII text editor in the DOS environment (via various

<sup>9</sup> [<http://www.cuneiform.de/fortschritte/webviewer.html>].

adaptable programmes such as Signum and Word Perfect) to Word for Windows, which still did not guarantee compatibility and a smooth file exchange. For a project like HPM, however, it was crucial to store and provide its data in a documented, open-source format and thus be independent not only of proprietary software, but also of specific computer platforms.

After twenty years of such struggles, the dynamics of the World Wide Web created an adaptation and standardization pressure that paved the way for creating platforms like HPM. When HPM went online, most of the transliteration letters for ancient Near Eastern languages were already available and could be displayed with some universal fonts (e.g., Arial Unicode, Code2000), which had to be installed in the operating system. For others (e.g., half brackets) similar characters had to be used and adapted in the Semiramis Unicode font that was developed for HPM. In 2008, in the course of an extension of Unicode, the half brackets and some ancient Egyptian transcription characters were given their own code point, prompting an adaptation in HPM. Today the encoding of the characters can be regarded as stable in the long term.

Based on Linux Libertine and Linux Biolinum, HPM created its own new set of fonts, which will be used for publications in print and online. Semiramis Unicode exists in a version 3, but is deprecated and will no longer be updated. In addition, S. Vanséveren provided a Unicode font for Hittite cuneiform and G. Anders for Luvian hieroglyphs.

Above the level of single characters, every document has to exist in an intelligible and well-documented format: as a kind of XML to describe textual data, as TIFF for photographs, SVG for drawings, or PLY for 3D data. For each of these data formats plenty of software is available; they are well documented and will permit the creation of new software in case this should be required. At HPM, this does not necessarily imply that all data will be made available to external users in these formats, mainly due to copyright issues.

### **13.5.3 Continuity Online: Development and Experiences**

The functionality and presentation of web pages with HTML has remained largely stable. The diversity of displays and media in the web has led to a growing awareness of the separation of content and form through multiple media. HPM is currently undergoing a revision in order to remove display-oriented elements deprecated by HTML 5.

The development of CSS was less consistent and the implementation of standards in web browsers sluggish. A careful use of HTML and CSS has low maintenance requirements and ensures sustainability. The implementation of the new media features of HTML 5 will further increase HPM's longevity.

Most websites on HPM also contain a dynamic component, e.g., the output of a database. The choice of technology for dynamic websites must be based on long-term

functionality. The programme code should be designed in a simple and modular way with little nesting in order to facilitate the acquisition of maintenance competence. In the humanities, but especially in “small disciplines” like Hittitology, highly qualified maintenance specialists may not be available or affordable at all times.

The uniformity and popularity of Content Management Systems (CMS) on the Internet, and the complexity of their numerous features, produce security vulnerabilities that are a popular gateway for hacker attacks. The use of a CMS therefore requires permanent and professional system management. The effort required for the integration of additional functions into a CMS causes further costs. The dependency on one CMS also limits flexibility and may impede a move from one host institution to another. For all these reasons, HPM does not use a CMS. The strength of a CMS as a multi-user system that ensures front-end homogeneity is less significant for HPM, where the visibility of the *individual* researchers and projects would even be undermined by a corporate approach to web design.

HPM uses PHP in conjunction with an Apache server and MySQL database. This is the standard configuration on university servers and thus ensures efficient hardware provisioning and data backup. With the introduction of PHP 7, revisions have now become inevitable. The outdated POSIX engine for regular expressions will be replaced by PCRE; also the deprecated connection to the MySQL database has to be replaced. These changes produce some uncertainty with regard to long-term maintenance.

Another instability concerns the presentation of media. The first photo viewer of HPM was written in Java. It had to be installed in web browsers, causing countless help requests from the HPM user community; also the performance of Java (speed) was unsatisfactory. The Flash viewer as its successor does a good job, but is notorious for its security flaws and will now be replaced by an HTML 5 viewer based on the HTML canvas implemented in the browsers. This promises long-term stability. JavaScript has also made great strides in the standardization process, so that a careful use of JavaScript can be considered sustainable. JavaScript is now so powerful that even 3D objects can be displayed in HTML 5 with its libraries.

#### **13.5.4 Tools for Scholars, not Scholars for Tools**

One of HPM’s basic goals is to collaborate with scholars without requiring them to leave their accustomed digital workflows and work environment. HPM offers simple avenues for accessing and processing data and manuscripts, which may have been created for a wide variety of purposes. In preparing materials for publication on HPM (manual), routine work should be avoided, and the revision of existing data kept to a minimum. For example, the manuscript of *Groddeks Liste* consists of several thousand pages arranged consistently by the author. By following a few rules, this document can be converted into a searchable online database, and an XML version is created automatically.

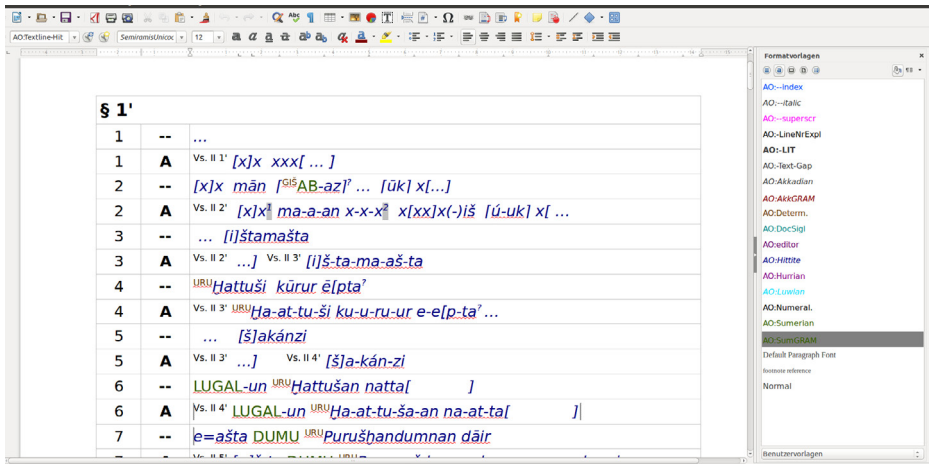


Figure 13.3: Hittite text with mark-up in LibreOffice

In other cases with more complicated structures, meaningfully named styles are added into the manuscript. In OpenOffice/LibreOffice this can be done in Fill Format Mode comparable to using a text marker on paper (Figure 13.3). The underlying XML structure can then be evaluated automatically, and the document can be converted into any XML format. In principle, the same method is used for creating the text editions on HPM. In the early years, HPM had developed a proprietary XML editor. For various applications, however, OpenOffice, which introduced XML into the Office sector, proved to be the best solution. It allows the quick definition of tags, records revisions and permits the use of foot- and endnotes, all as XML that can be processed and read in any text programme (Figure 13.4).

**harganu-**

Hutter Braunsar 1989a, 204 (militär. Kontext).

**harki-, \*harkant-**

Giorgadze 1988a, 69ff.

**harnai- (Verbum u. Subst.)**

s.u. **hu/arnai-**

**nindaharnantašši-**

Hutter 1988a, 57f. (s. aber auch Melchert CLL, s.v. \*harnant(i)-; ibid, auch heth. Lehnw. harnanta-).

**(aš)harnašalla-**

Siegelová 1986a, 67, Anm. 8 (= akk. 𒀭𒀭𒀭𒀭𒀭𒀭𒀭𒀭, 506, Anm. 1.

**harnau-/harnu-**

Pringle 1985a, 658 (zu SAL harnauwaš).

Figure 13.4: An example for the reuse of an older manuscript as database by tagging it with styles

To facilitate and speed up work, Müller developed the programme Simtex for the digitization of larger bodies of texts. This simple input method generates XML from text files automatically, which is then further processed as an OpenDocument.

### 13.5.5 Connecting Data

HPM makes data collections and tools for research available online. A particular advantage of the digital medium is the possibility of linking data sets, which often results in new findings. As far as possible, information should be collected only once and then be linked to other information, whether it is word forms, datings, locations, place names, persons, or inscribed objects. This does not exclude competing projects, but the goal is to preserve existing data collections and develop them further in a continuous, joint effort.

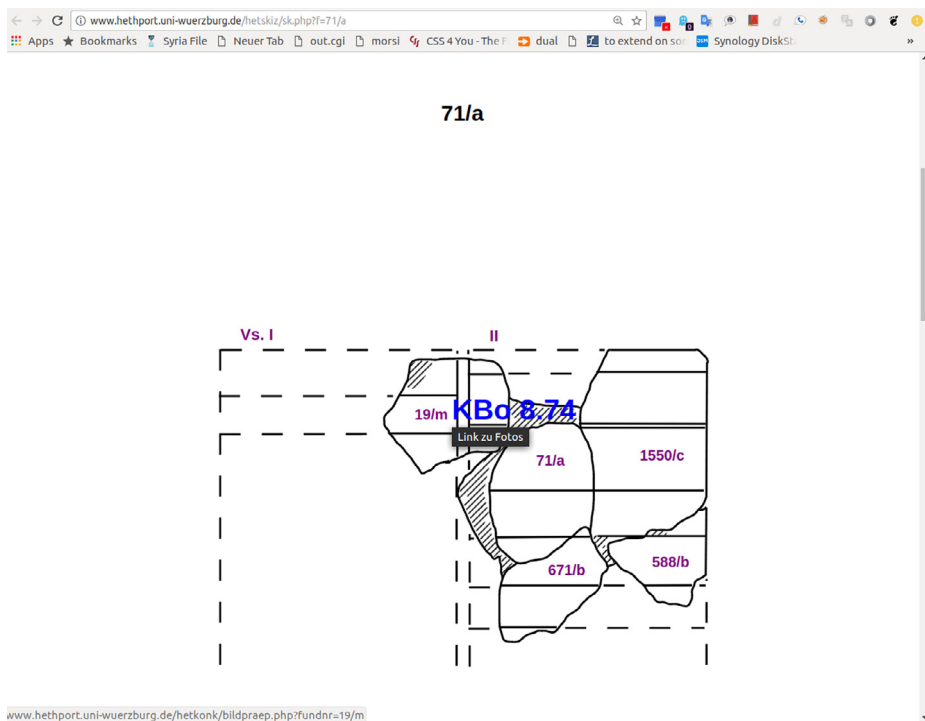
The data must be arranged in such a way that any correction has to be made in only one place, and is then available at all points of reference automatically. This seems trivial for an Internet project, but does indeed pose a challenge due to the heterogeneity of the individual projects that collaborate with HPM.

The principle of collecting any given information in only one place has further important implications, not least with regard to information that has to be retrieved multiple times. Thus a text edition should be limited to the text and not contain a lexical, morphological and syntactic annotation. An integrated annotation not only requires the repeated annotation of the same word form; it also impedes future changes. If the analytic annotation is separate from and, at the same time, linked to the text editions, the work process is more efficient, and it is easier to preserve consistency over time.<sup>10</sup>

An example of the efficient arrangement of information is the *Joinskizzen* component, which provides a documentation of the reconstruction of fragmented cuneiform tablets in drawings. As a stable reference, the join sketches include the inventory number of the individual fragments in the drawing. But the citation of fragments in the literature usually is by publication number. The join sketches were originally drawn in ink and published in print. Later they were digitized for the online version. The drawings were transferred to OpenOffice Draw, which uses a vector graphics format. This is automatically processed and the individual parts can be addressed. When the user is loading a join sketch, a link to the relevant photos is automatically generated. The publication number is automatically queried in the *Konkordanz* and appears on mouse over (Figure 13.5).

---

<sup>10</sup> The fact that in the Hurrian language no fewer than five new nominal cases have been discovered over the last thirty years may serve as a warning against hard-coding every analysis in the text edition.



**Figure 13.5:** On mouse-over the SVG join sketch displays the publication number of the fragment and a link to the relevant photos

In the future, HPM intends to extend the offer of such automatic links. Sometimes there are, however, compatibility issues: for example the *Konkordanz* and *Groddeks Liste* do not always follow the same bibliographical standards for referencing cuneiform manuscripts by publication.

### 13.6 Outlook: Expansion, Connectivity, Sustainability

All components of HPM – catalogues, bibliographies, onomastica, text editions, media archives, and services – are engaged in a continuous process of further development and expansion. The single most important new component currently under preparation is the *Thesaurus Linguarum Hethaeorum digitalis* (TLH<sup>dig</sup>), a tool that is not conceived as a dictionary, but as a basic, searchable database of all cuneiform manuscripts from Hittite tablet collections in transliteration. In this context, HPM is planning to develop a “creator interface” that will allow users to actively contribute newly discovered texts to the growing TLH<sup>dig</sup>. This will add a new dimension to HPM as a truly collaborative digital infrastructure.

External connectivity is an important challenge for the future, especially when more Hittitological research publications will become available in a digital format. This particularly concerns the Hittitological philological dictionaries, which currently have only a limited or no digital presence, and the various databases of excavations in the Anatolian cultural area, only some of which are accessible online. Also the creation of dynamic interconnections (rather than static links) between the various digital corpora of cuneiform texts (see fn. 2) is an important task for the future.

Expansion and connectivity must be underpinned by sustainability in order to ensure the long-term availability of HPM. The structure of the existing web pages must be further developed to allow easy maintenance. Not only the data themselves should be present in documents (XML) that are self-explanatory, but also the programme logic should be stored in forms that allow automatic reconfiguration. This is likely to involve an increased use of XML technologies (with XSL and XML databases), but such changes should be approached with the necessary caution and employ only common, wide-spread and proven technologies.

## Bibliography

- Laroche, E. (1971). *Catalogue des textes hittites*. Paris: Klincksieck.
- Müller, G.G.W. & Wilhelm, G. (2015). *Informationsinfrastruktur für digitale Publikation keilschriftlicher Staatsverträge der Hethiter und für darauf bezogene netzbasierte Forschungskoooperation*. Retrieved from [<http://www.hethiter.net/HPM/hpm.php?p=projektDFGINfrastruktur>], 2017/10/31.
- de Roos, J. (2007). Review of: S. Košak, Konkordanz der hethitischen Keilschrifttafeln. *Bibliotheca Orientalis*, 64, 187–188.
- Schwemer, D. (2017). Hittitology at the Centennial. Current Trends and Future Directions in the Study of Hittite Culture. In M. Doğan-Alparslan, A. Schachner, & M. Alparslan (Eds.), *The Discovery of an Anatolian Empire. A Colloquium to Commemorate the 100th Anniversary of the Decipherment of the Hittite Language* (pp. 295–303). Istanbul: Türk Eskiçağ Bilimleri Enstitüsü.
- Wilhelm, G. (2008). Die Edition der Keilschrifttafeln aus Boğazköy und das Projekt “Hethitische Forschungen” der Akademie der Wissenschaften und der Literatur, Mainz. In G. Wilhelm (Ed.), *Ḫattuša – Boğazköy. Das Hethiterreich im Spannungsfeld des Alten Orients. 6. Internationales Colloquium der Deutschen Orient-Gesellschaft, 22.–24. März 2006, Würzburg* (Colloquien der Deutschen Orient-Gesellschaft 6) (pp. 73–86). Wiesbaden: Harrassowitz.
- Wilhelm, G. (2013). Das Hethitologie Portal Mainz. In I. Hafemann (Ed.), *Perspektiven einer corpusbasierten historischen Linguistik und Philologie. Internationale Tagung des Akademievorhabens „Altägyptisches Wörterbuch“ an der Berlin-Brandenburgischen Akademie der Wissenschaften, 12.–13. Dezember 2011* (pp. 155–69). Berlin: Berlin-Brandenburgische Akademie der Wissenschaften.
- Wilhelm, G. (2015). Das Projekt “Hethitische Forschungen” der Akademie der Wissenschaften und der Literatur, Mainz. Retrieved from [<http://www.hethiter.net/HPM/hpm.php?p=projektHethForsch>], 2017/10/31.

Nadia Cannata

## 14 EDV – Italian Medieval Epigraphy in the Vernacular Some Editorial Problems Discussed

**Abstract:** EDV (Epigraphic Database Vernacular) is a database collecting the vernacular inscriptions produced in Italy from the late Medieval to the Early Modern Age, and is a part of the EAGLE and IDEA projects. The present contribution illustrates the criteria used for the description and indexing of all inscriptions that record public script in language(s) other than Latin. The material is very varied as regards language, script, provenance, support and function. The author discusses briefly the editorial criteria that may prove most appropriate for its publication.

**Keywords:** medieval epigraphy, textual criticism, Romance linguistics, digital humanities, palaeography

### 14.1 The Corpus

EDV is a new database recording the corpus of all vernacular inscriptions that were produced in Italy from the middle of the 9<sup>th</sup> century to the year 1500 CE, provided they were meant to be displayed publicly and are still extant. The aim of the study – which has been progressing since 2011 – is to collect documentary evidence of the uses of language(s) other than Latin in public script in late Medieval and Early Modern Italy. As I write, EDV contains over 530 items, and new entries are constantly being added (albeit at a slowing pace now). We intend to publish the complete catalogue both through a website currently under construction<sup>1</sup> and in book form.<sup>2</sup>

---

1 [www.edvcorpus.com/wp/]. The corpus is to be hosted on the EAGLE platform (Europeana network of Ancient Greek and Latin), a best-practice network co-funded by the European Commission, under its Information and Communication Technologies Policy Support Programme, and is now part of IDEA. International Digital Epigraphy Association [http://www.idea-association.eu] (see Chapter 17 in this volume, and Orlandi et al., 2017).

2 The work was started as the subject of the MA and doctoral dissertations of Drs Luna Cacchioli and Alessandra Tiburzi, supervised by me at the University of Roma “La Sapienza”. The first results have been published in three different contributions: Cacchioli & Tiburzi (2014, 2015); Cacchioli, Cannata, & Tiburzi (2016), and a book is in preparation (Cacchioli, Cannata, & Tiburzi, 2019).

---

Nadia Cannata, Università di Roma, “La Sapienza”



© 2018 Nadia Cannata

This work is licensed under the Creative Commons Attribution 4.0 International license (CC BY 4.0)



The number of inscriptions so far identified and recorded widely exceeds our previous knowledge of the extent to which the language of the *illiterates* (i.e. those who did not know Latin) was used in public contexts. EDV collects them as a category for the first time. Therefore, even though quite a few of the inscriptions have received scholarly attention and are known (often very well known) to the scholarly community, it has not been possible – before all the data were collected and made available – to look at the historical phenomenon of public script in Early Modern Italy, produced in languages other than Latin in its entirety. In this respect, EDV constitutes material for a new discipline – medieval and early modern vernacular epigraphy – the study of which may be of interest not only to epigraphists and linguists (both philologists in general, and Romance philologists in particular), but also to scholars engaged in fields of enquiry as diverse as culture history and anthropology, palaeography, history of art and architecture.

## 14.2 The Background

In 1967, Augusto Campana's seminal article advocated, for the first time, the need to establish an epigraphic scholarship concerning itself with the study of early modern inscriptions (Campana, 1967). He argued that an inscription needs to be investigated and interpreted through the joint cooperation of palaeographers, art historians and linguists, since only such cooperation would allow for it to be fully understood in all its components: text, script and monument. This, of course, applies to inscriptions produced in any language. Nearly four decades later, the first systematic catalogue of medieval inscriptions was launched: IMAI – *Inscriptiones Medii Aevii Italiae (saec. VI–XII)*, a series that aims to catalogue all inscriptions produced in Italy within that chronological span. It is organized according to Italian administrative regions (so far the volumes *Lazio – Viterbo*, *Umbria – Terni* and *Veneto* were published), and offers the text of the inscriptions (in both diplomatic and critical editions), a photographic reproduction of the pieces published, accompanied by a detailed palaeographic analysis of the script(s) used (Cimarra, Condello, Miglio et al., 2002; Guerrini, 2010; De Rubeis, 2011).

The interest in vernacular epigraphy has flourished somewhat later, but it yielded its first results at a quicker pace (Petrucci 1985, 1986, 1988). In 1995, Claudio Ciociola organized an exhibition and conference, *Visibile parlare* at the University of Cassino. The proceedings of the conference were published in a volume that effectively signalled the start of a dedicated scholarly interest for this body of texts. Perhaps bearing in mind Campana's remarks, the volume is arranged into three sections devoted, respectively, to Palaeography, Language History and Textual Criticism, and Iconography (Ciociola, 1997).

Some pioneering work has also been carried out on the use of script in Renaissance art. The first, and perhaps major such contribution is Dario Covi's 1958

PhD dissertation, now published as *The inscription in Fifteenth Century Florentine Painting* (Covi, 1986). Covi's work is complemented by A. Dietl's *Die Sprache der Signatur* (Dietl, 2008), and by the periodical *Opera Nomina Historiae*, launched by the late Maria Monica Donato at the Scuola Normale Superiore in Pisa.<sup>3</sup>

During the past twenty years many new projects were undertaken, and have revealed a remarkable treasure of texts, mostly the result of interest in local history and linguistics. Particular attention has been devoted to Rome and the Lazio (Sabatini, Raffaelli, & D'Achille, 1987; Sabatini, 1996; Tedeschi, 2012, 2014), Venice and the Veneto (Tomasin, 2001, 2004, 2012a, 2012b, 2013; Di Lenardo, 2014; Benucci, 2015; Ferguson, 2015), as well as to the earliest examples in Tuscany and elsewhere. A recent volume (Petrucci, 2010) catalogues all vernacular inscriptions produced up to the 13<sup>th</sup> century, and other similar cataloguing initiatives are also being undertaken outside of Italy.<sup>4</sup> Therefore, we have a significant body of texts and templates to guide our criteria in setting up EDV, which aims at generating a comprehensive catalogue of Italian vernacular inscriptions, similar to IMAI, but with certain differences that may be worth discussing (Geymonat, 2014).

### 14.3 History, Geography, Forms and Functions

For historians, it is a known fact that the nature of the documents they are studying should shape the form through which such material is published, edited and circulated. The main feature of the corpus contained in EDV lies in its complex variety, in terms of time, geography, types, form, function, language and script.

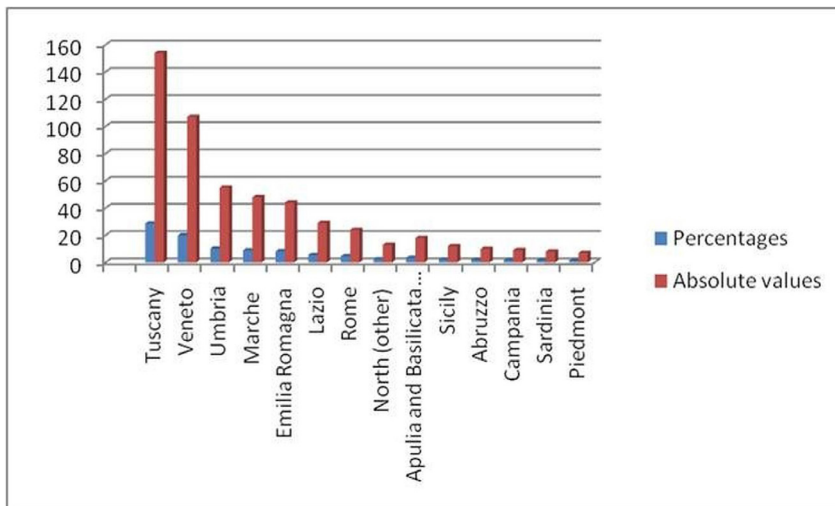
The inscriptions were written in different vernaculars, in Catalan, and in Old French. Some are informal notices, other are epigraphs solemnly celebrating patronage. Many were inscribed in stone or engraved on metal, some are casually scratched on plaster. Many are painted on wood or canvas: sometimes solemnly displayed, sometimes disguised in the picture, or else discreetly placed to indicate the authorship of a painting or the biblical source of a scene. Those cut into stone were in most, if not all cases, not created by the authors of the text they bear. Others are written or scratched by whomever devised the message they convey. In both cases they show a degree of skill in using script, which may range from the barely literate to the highly professional.

<sup>3</sup> [<http://onh.giornale.sns.it/>].

<sup>4</sup> I am thinking of the project entitled *Écritures Exposées. Discours, matérialité et usages* jointly coordinated by the École des hautes études hispaniques et ibériques (Casa de Velázquez, Madrid), the Grupo de Investigación "Lectura, escritura, alfabetización" (LEA), Seminario Interdisciplinar de Estudios sobre Cultura Escrita (SIECE) (Universidad de Alcalá) and the Centre d'Etude des Littératures et Langues Anciennes et Modernes (CELLAM), Groupe de recherche sur culture écrite et société (GRECES) (Université Rennes 2).

All these features should be covered thoroughly for the database to be of any use to scholars, and organized according to categories that design a taxonomy that has historical significance.<sup>5</sup> Let us consider, for example, the geographic distribution of the inscriptions. Reasons of practicality suggest the use, as general categories for localization, of the administrative regions of modern Italy (Lombardia, Veneto, Tuscany, Apulia, Sicily and so on). Some of those – Tuscany or Sicily to name only two – constitute a monument to Italian history, and have existed since the late middle ages with that very designation. Dante, in the *Comedia*, is addressed by his fellow Florentine citizen Farinata as “O Tosco” (Inf. X, 22). Others, however, would have been unknown at the time when the documents were produced. For example, Lazio did not exist before 1927. Dante used the term to refer to “Italy” as the land where Latins live (*De Vulgari Eloquentia*, I, *passim*).

In absolute terms Tuscany and the Veneto – accounting, respectively, for 154 and 107 inscriptions – house the highest number of inscriptions per area, and between them they cover nearly half (49%) of the corpus, as shown in the following chart indicating the number of inscriptions per region and the percentage the region occupies in the sample (Figure 14.1).

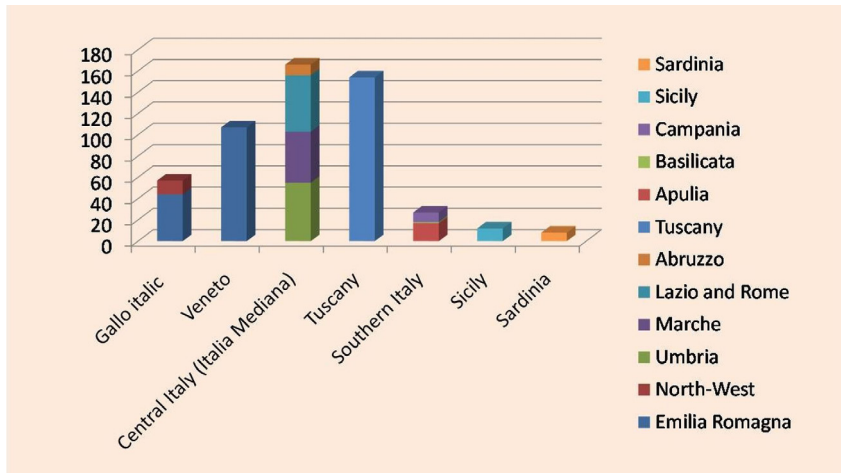


**Figure 14.1:** Number of inscriptions per region (in red) and the percentage the region occupies in the sample (in blue)

<sup>5</sup> Similar issues are being addressed by many other projects that deal with the publication, in both paper and digital form, of epigraphic materials of a multilingual nature. See, for example the Hesperia project (discussed in Chapter 3 in this volume), the OCIANA project (Chapter 8) and the I.Sicily project (Chapter 19).

If, however, we were to classify our data according to the vernaculars used, it would be more useful to adopt a different grid, and distinguish between Northern Italy where Gallo-italic vernaculars were spoken, Friuli and the Veneto, Central Italy excluding Tuscany (the so-called *Italia mediana* which includes part of the Abruzzi), Tuscany, Southern Italy (Southern Abruzzi, Campania, Basilicata, Northern Apulia and Calabria), Extreme South (southern Apulia and Calabria, Sicily), and Sardinia.

To Northern Italy (from Valle d'Aosta down to, and including, Emilia Romagna) belong 171 inscriptions (32%), Central Italy accounts for 156 items (29%) (Figure 14.2).



**Figure 14.2:** Number of inscriptions per region (according to the vernacular used)

If we were to include Tuscany where it geographically belongs, we would see a very different picture; one that shows Central Italy as the area where the vernacular was most widely employed in public life, and where it replaced Latin in many of its functions. Conversely, only a mere 11% of the inscriptions are attributed to Southern Italy, which – one might be inclined to think – remained more aligned with tradition (Figure 14.3).

But was it really? The imbalance demonstrated by the data is also due (maybe largely due) to the greater documentation available for Tuscany and Venice, thanks to the position they occupy in Italian history and culture. More scholarly attention naturally results in more documentation being available, which in turn could cause their standing out from the rest of the sample, perhaps more so than the facts would allow. The eye of the beholder alters the picture that is seen, and never more acutely is this the case than when we deal with the large volume of data that digital humanities make available. We need, therefore, to allow for the data to be considered under different headings and perspectives, and to be interrogated with the greatest possible flexibility.

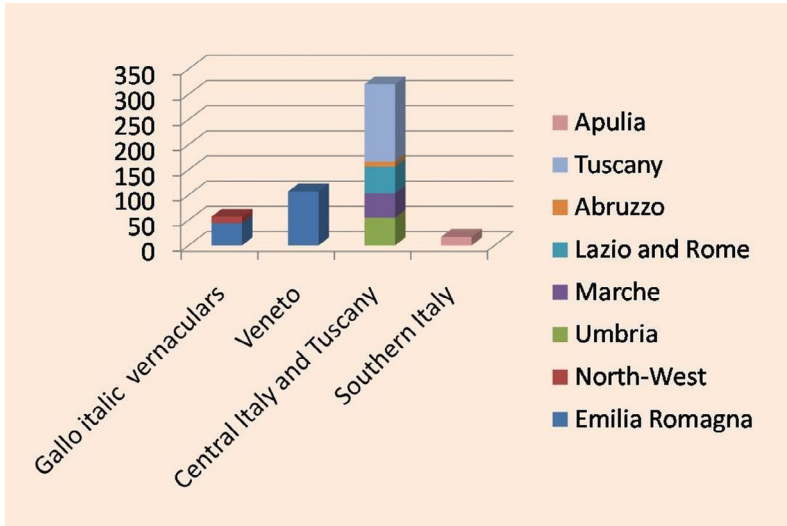


Figure 14.3: Use of vernacular in Central and Southern Italy

## 14.4 How are the Data Organized

Public script usually records the language dominant in a community at any given time. It constitutes one among the noblest forms of writing (Petrucci in Ciociola, 1997), because of its solemn and formal nature, which is usually appropriate to the dignity and importance of the message conveyed (be it the commemoration of the dead, of an event, or the issuing of a law). Public script also documents the relationship between orality and written records i.e. the language(s) in use in speech and in written documentation. They do not necessarily coincide, and literature often uses a language that may be very different, or a different language altogether, from the spoken language, especially in Medieval Italy. Inscriptions tend, however, to document a language that approaches more closely what was in use in the community. Inscriptions were produced to convey messages addressed to an entire group and – in order to be effective and to fulfil their function – they should have been written in a language understood by most, if not all. Often, since they were preserved to this day, they appear to have been valued by that community across the centuries as monuments to a shared past and shared identity.

The inscriptions included in the catalogue span nearly six-hundred years of history, and were produced across the whole peninsula in many different languages and scripts for a different array of functions. They were also engraved, painted or scratched onto a variety of different writing surfaces ranging from stone to plaster, wood, cloth, metal (gold, silver, bronze, iron), terracotta and ivory. All this information needs to be recorded and searchable within the online database, as well as in print.

For the purpose of this chapter we need to distinguish between the General Catalogue of the inscriptions, which will be openly accessible to all visitors of the EDV website, and the materials that will be published in print only. The website and book cover different functions and will therefore be used for different forms of publication.

We believe that the material we have identified and studied needs to be presented within a critical framework that online browsing does not allow, because it would make its consultation highly impractical.

In addition, we would like to address explicitly the issues of sustainability and durability. The high cost of maintaining digital records available over time is an issue for all researchers engaged in digital epigraphy projects.<sup>6</sup> In our opinion it is advisable to provide a paper edition of the database in traditional book form, which is best suited to accommodate the complexity of the data (text in critical edition, linguistic analysis and paraphrase, all historical information on the building or painting hosting the inscription). The General Catalogue, available online, is organised under eight headings as follows:

1. ORIGIN (Region of production according to modern Italian administrative regions)
2. DATE (century, half century, quarter, year – as available)
3. CURRENT LOCATION (Site, Church, Museum etc.)
4. PLACE OF PRODUCTION (City, Town, Village)
  - (a) Linguistic area
5. IDENTIFICATION (e.g. General title, e.g. *Iscrizione di Commodilla, Lauda di Vanzone*)
6. WRITING SURFACE
  - (a) Bronze
  - (b) Canvas
  - (c) Copper
  - (d) Fabric
  - (e) Gold
  - (f) Iron
  - (g) Ivory
  - (h) Mosaic
  - (i) Plaster
  - (j) Silver
  - (k) Stone
  - (l) Terracotta
  - (m) Wood

---

<sup>6</sup> An issue addressed by the creation of the IDEA network (see above, note 1).

7. TYPE (General Category according to the nature of the inscription, articulated in sub-categories, according to function, as appropriate):

(a) Public Notices

1. Memorials of major events (floods, pestilence, coronations etc.)
2. Patronage
3. Rulings (edicts, laws etc.)

(b) Captions

1. Admonitions (proverbs, adages, moral statements etc.)
2. Narrative captions in paintings
3. Artists' signatures

(c) Funerary inscriptions<sup>7</sup>

(d) Inscriptions on objects of everyday use

(e) Graffiti and other extemporary notes

8. SCRIPT

(a) Gothic

(b) Capital

(c) Mixed scripts (elementary level)<sup>8</sup>

9. ICONOGRAPHY

10. PHOTOGRAPHIC REPRODUCTION (Yes/No)

11. BIBLIOGRAPHY

The website also hosts a blog and an area for readers to give notice of any new items, or report mistakes or missing information, and anything else that may be of interest in relation to the corpus.

The printed edition of the corpus will include: a brief description of each item; the context in and for which it was produced; a summary of its content; the complete text of each inscription, both in diplomatic and critical edition;<sup>9</sup> a critical apparatus; a detailed linguistic commentary of the text (phonology, morphology, syntax and lexicon); and a palaeographical commentary and bibliography (Cacchioli, Cannata, & Tiburzi, 2016).

---

<sup>7</sup> Funerary inscriptions should be a sub-category of public inscriptions, but given their numerosity and unique nature we have kept them separate.

<sup>8</sup> Scripts which were executed at a very low level of skill and cannot therefore easily be classified are recorded as "elementary".

<sup>9</sup> The metadata will be imported directly into the database. For each inscription an XML file containing the text elements encoded according to EpiDoc will be created (issues related to the encoding are extensively explained in Cacchioli, Cannata, & Tiburzi, 2016). The XML files published in the database (only a small selection) will be available for downloading. At the moment we will only tag inscriptions for the study purpose of the research group, at least for the time being.

The template of each entry is as follows:

ORIGIN (The data are arranged according to modern administrative regions)  
 TITLE  
 LOCATION  
 DATE  
 MATERIAL  
 MEASUREMENTS  
 TYPE  
 FUNCTION  
 NOTES  
 DIPLOMATIC EDITION  
 CRITICAL EDITION  
 LINGUISTIC ANALYSIS (phonology, morphology and syntax, lexicon)  
 SCRIPT  
 BIBLIOGRAPHY  
 PHOTO

Here is an example:

LAZIO  
*Telamone erratico*<sup>10</sup>  
 Ferentino (FR), Chiesa dei santi Giovanni e Paolo (Duomo)  
 1220-1230  
 Stone  
 TYPE  
 Caption  
 FUNCTION  
 Narrative  
 NOTES

The short text is engraved at the basis of the stone basin supported by the telamone erratico. It is preserved in the Church of the Santi Giovanni e Paolo in Ferentino. It constitutes a lamentation about the weight of the stone.

V  
 P  
 E  
 S  
 A

---

<sup>10</sup> The author of the record is Luna Cacchioli.



U[h],/p/e/s/a!

SCRIPT:  
Capital.

LINGUISTIC ANALYSIS:

The inscription documents the first known use of both ‘uh’ and the verb ‘pesare’. Deli, dates *u[h]* generically as before 1492 and the verb ‘pesare’ before 1320.<sup>11</sup>

*Editions:* D’ACHILLE 2012, p. 112.

*Photos:* D’ACHILLE A. M. 2012, fig. 32, p. 92.



## 14.5 Conclusion

In his *Sermones*, Augustine claimed that walls might sometimes function as open books. Indeed, throughout the early modern period, texts and images were often

---

<sup>11</sup> DELI (see Cortelazzo & Zolli, 1979–1988), s.v. *uh* and *pesare*.

used to enrich the walls of churches, private homes, public palaces and other seats of power. As literacy spread, and with it the public use of script, moral admonitions, proverbs, captions and signatures in paintings, mementoes of patronage or of some catastrophic event, laws and edicts, as well as funerary inscriptions, all appeared with increasing frequency. In Italy, the use of languages other than Latin in public life, and as verbal complements to artistic representation, also intensified over time at an increasing pace, and became rather dominant during the Quattrocento.

The sheer wealth of the material uncovered will certainly help understand how, to what extent and why, languages other than Latin were used as a complement to visual arts and architecture in Italy in the early modern period. It will also provide fresh material for the study of the relation established in time and place between language and the aesthetics of an artefact and the role that issues of verbal communication played within artistic representation. The material might also help document if, and how, writing was used as an adornment in contemporary art, as well as provide very useful information relating to the sociolinguistics of early Italian (when and why was the modern language and its varieties used in lieu of Latin and for what purposes), and the spread of a common language in Italy well before the Cinquecento, which is when we conventionally date the birth of the national language.

The template adopted aims to be able to organize data in a useful manner for the purposes listed above, by providing a flexible and historically accurate tool for the study of the materials made available to the wider community of scholars. It aims at catering for the needs of all historians, regardless of their field of specialization, and it is expected it will prove to be flexible and open to correction, resilient, and most of all, durable in time.

## Bibliography

- Benucci, F. (Ed.). (2015). *Corpus dell'Epigrafia medievale di Padova. 1: Le iscrizioni medievali dei Musei civici di Padova. Museo di arte medievale e moderna*. Sommacampagna: Cierre edizioni.
- Cacchioli, L. & Tiburzi, A. (2014). Lingua e forme dell'epigrafia medievale in volgare. *Studi Romanzi, n.s. X*, 311–152.
- Cacchioli, L. & Tiburzi, A. (2015). Contributi e fonti per lo studio del volgare esposto in Italia, *Critica del Testo, XVIII(2)*, 103–138.
- Cacchioli, L., Cannata N., & Tiburzi, A. (2016). EDV – *Italian Medieval Epigraphy in the Vernacular (IX-XV c). A New Database*. In A.E. Felle & A. Rocco (Eds.), *Off the Beaten Track: Epigraphy at the Borders. Proceedings of 6th EAGLE International Event (24-25 September 2015, Bari, Italy)*. Oxford: Archaeopress (pp. 91–129). Retrieved from [http://www.archaeopress.com/ArchaeopressShop/Public/download.asp?id={E7B2AAC6-9986-4C41-9842-6AA93BE7ACD9}], 2017/11/29.

- Cacchioli, L., Cannata, N., & Tiburzi, A. (2019, forthcoming). *Il volgare esposto in Italia, secc. IX-XV. Lingue, scritture, funzioni* (con una nota paleografica di Maddalena Signorini). Roma: Bagatto Libri.
- Campana, A. (1967). Tutela dei beni epigrafici. *Per la salvezza dei beni culturali in Italia. Atti e documenti della Commissione d'indagine per la tutela e la valorizzazione del patrimonio storico, archeologico, artistico e del paesaggio* (Vol.2) (pp. 539–547). Roma: Casa Editrice Colombo.
- Cimarra, L., Condello, E., Miglio, L., Signorini, M., Supino, P., & Tedeschi, C. (2002). *Inscriptiones Medii Aevi Italiae (Saec. VI-XII). Vol. 1. LAZIO-Viterbo*. Spoleto: Centro Italiano di Studi sull'Alto Medioevo.
- Ciociola, C. (1989). Visibile parlare: agenda. *Rivista di letteratura italiana*, VII, 9–77.
- Ciociola, C. (Ed.). (1997). *Visibile parlare: le scritture esposte nei volgari italiani dal Medioevo al Rinascimento. Atti del Convegno internazionale di studi (Cassino-Montecassino 26-28 ottobre 1992)*. Napoli: Edizioni scientifiche italiane.
- Cortelazzo, M. & Zolli, P. (1979–1988). *DELLI. Dizionario etimologico della lingua italiana*. Bologna: Zanichelli.
- Covi, D. (1986). *The Inscription in Fifteenth Century Florentine Painting*. New-York-London: Garland.
- D'Achille, P. (2012). *Parole: al muro e in scena. L'italiano esposto e rappresentato*. Firenze: Franco Cesati Editore.
- De Rubeis, F. (2011). *Inscriptiones Medii Aevi Italiae (Saec. VI-XII). Vol. 3. VENETO - Treviso, Vicenza, Belluno*. Spoleto: Centro Italiano di Studi sull'Alto Medioevo.
- Dietl, A. (2008). *Die Sprache der Signatur. Die mittelalterlichen Künstlerinschriften Italiens*. München-Berlin: Deutscher Kunstverlag.
- Di Lenardo, L. (2014). *La collezione epigrafica del Seminario patriarcale di Venezia. Catalogo (secoli XII-XV)*. Venezia: Marcianum Press.
- Ferguson, R. (2015). *Le iscrizioni in antico volgare delle confraternite laiche veneziane: edizione e commento*. Venezia: Marcianum Press.
- Geymonat, F. (2014). Scritture esposte. In G. Antonelli, M. Motolese, & L. Tomasin (Eds.), *Storia dell'Italiano scritto. Vol. 3. Italiano dell'uso* (pp. 57–100). Roma: Carocci.
- Guerrini, P. (2010). *Inscriptiones Medii Aevi Italiae (Saec. VI-XII), UMBRIA ~ Terni*. Spoleto: Centro Italiano di Studi sull'Alto Medioevo.
- Orlandi, S., Santucci, R., Mambrini, F., & Liuzzo, P.M. (Eds.). (2017). *Digital and Traditional Epigraphy in Context. Proceedings of the EAGLE 2016 International Conference*. Roma: Sapienza Università Editrice. doi: 10.13133/978-88-9377-021-7
- Petrucci, A. (1985). Potere, spazi urbani, scritture esposte. Proposte ed esempi. In *Culture et idéologie dans la genèse de l'état moderne. Actes de la table ronde organisée par le Centre National de la recherche scientifique et l'École française de Rome (Rome, 15-17 octobre 1984)* (pp. 85–97). Rome: École française de Rome.
- Petrucci, A. (1986). *La scrittura: ideologia e rappresentazione*. Torino: Einaudi.
- Petrucci, A. (1988). Storia e geografia delle culture scritte (dal secolo XI al secolo XVII). In A. Asor Rosa (Ed.), *Letteratura italiana. Storia e geografia. II, L'età moderna* (pp. 1193–1292). Torino: Einaudi.
- Petrucci, L. (2010). *Alle origini dell'epigrafia volgare. Iscrizioni italiane e romanze fino al 1275*. Pisa: Edizioni Plus.
- Sabatini, F. (1996). Voci nella pietra dall'Italia mediana. Analisi di un campione e proposte per una tipologia delle iscrizioni in volgare. In F. Sabatini, *Italia linguistica delle origini, saggi editi dal 1956 al 1996*, collected by V. Coletti, R. Coluccia, P. D'Achille, N. De Blasi, L. Petrucci (pp. 569–625). Lecce: Argo.

- Sabatini, F., Raffaelli, S., & D'Achille, P. (1987). *Il volgare nelle chiese di Roma. Messaggi graffiti, dipinti e incisi dal IX al XVI secolo*. Roma: Bonacci.
- Tedeschi, C. (Ed.). (2012). *Graffiti Templari. Scritture e simboli medievali in una tomba etrusca di Tarquinia*. Rome: Viella.
- Tedeschi, C. (2014). I graffiti, una fonte scritta trascurata. In D. D. Bianconi (Ed.), *Storia della scrittura e altre storie* (pp. 363–381). Roma: Accademia Nazionale dei Lincei, Scienze e Lettere.
- Tomasin, L. (2001). La lapide veneziana di S. Gottardo a Piazzola sul Brenta (1384). *L'Italia Dialettale*, 63, 173–177.
- Tomasin, L. (2004). *Testi padovani del Trecento: edizione e commento linguistico*. Padova: Esedra.
- Tomasin, L. (2012). Epigrafi trecentesche in volgare nei dintorni di Venezia. *Lingua e stile*, 47(1), 23–44.
- Tomasin, L. (2012). Minima muralia. Esercizio di epigrafia medievale. *Vox romanica*, 71, pp. 1–12.
- Tomasin, L. (2013). Un'epigrafe ferrarese in volgare. *Quaderni veneti*, 2, 173–181.

Mark Depauw

## 15 Trismegistos: Optimizing Interoperability for Texts from the Ancient World

**Abstract:** Although its origins lie with the *Prosopographia Ptolemaica*, a project studying people who lived in Ptolemaic Egypt (332–30 BCE), Trismegistos has developed into an interdisciplinary platform for the study of the ancient world in general, from 800 BCE to 800 CE: texts, places, people, collections. Setting up this very divergent set of databases has only been possible through the availability of full text corpora, new digital processing techniques, and the “exponentiality” permitted by interconnectivity. By bringing everything together in a single environment, Trismegistos has facilitated quantitative studies of several phenomena, but this approach remains promising and will hopefully become more widespread. TM’s main aim, however, is interoperability through the spread of stable identifiers, as an instrument to build a Linked Open Data environment for the ancient world.

**Keywords:** interoperability, ancient world, metadata standards, Linked Open Data, stable identifiers

### 15.1 The Development of Trismegistos (Texts)

Trismegistos<sup>1</sup> is a relatively young project, launched in Cologne in 2006 in the framework of a Kovalevskaja Award of the Alexander von Humboldt Stiftung. Its roots, however, go back almost 80 years to Leuven, where it is also currently housed.

In 1937 Willy Peremans wrote *Vreemdelingen en Egyptenaren (Foreigners and Egyptians)*, which must be one of the last papyrological works for a scholarly audience in Dutch (Peremans, 1937). It would turn out to be programmatic for Ancient History at KU Leuven in several ways. In the first place, because Peremans realized that a thorough study of the relations between the newly arrived Greeks and the local populations would need a prosopography. After World War II, he therefore started the *Prosopographia Ptolemaica* [PP], which should become a list of all attested individuals living in Ptolemaic Egypt.

---

<sup>1</sup> TM [<http://www.trismegistos.org>].

---

Mark Depauw, KU Leuven



© 2018 Mark Depauw

This work is licensed under the Creative Commons Attribution 4.0 International license (CC BY 4.0)

Peremans' approach to this prosopography was, certainly for that time, remarkable: all-inclusive and interdisciplinary *avant la lettre*. Together with his assistant Van't Dack, he decided that documentary papyri and ostraca would be the core material, but information from epigraphic texts or literary sources would not be neglected either. As a Flemish nationalist, Peremans also insisted from the beginning that texts in the local Egyptian languages, Demotic and hieroglyphic would be included, even if he did not know them himself.

All this resulted in a series of printed volumes, each covering specific corporate categories, more or less following social hierarchy. The volumes were published between 1950 and 1968, with an index in 1975. However, the supplements to the early volumes, published in 1975 and 1981, already illustrated some infrastructural problems.<sup>2</sup> Within each category, people were ordered alphabetically by their name and assigned a number for ease of reference. Newly added individuals thus had to receive a letter in addition to a number, e.g. PP VIII 3844a.

This was obviously going to be a problem in the end, but fortunately, technology offered a solution in the form of the computer. The PP was an early adopter of this innovation, and started with the “computerisation of the documentation in a relational database” in the mid-eighties, around the time of Peremans' death in 1986 (Mooren, 2001). As it was never really a project with separate funding, much of this conversion work was carried out by assistants, and took quite some time. Paradoxically, the advent of the system of project funding in Leuven in the nineties did not really speed up the process, as Willy Clarysse and later Katelijn Vandorpe successfully applied for other projects such as the Leuven Homepage of Papyrus Collections [LHPC], the Leuven Database of Ancient Books [LDAB], the Fayum project, or the Archives project.<sup>3</sup>

Although the PP thus lied dormant in the early 2000's, the systematic data collection for it and for the other projects would turn out to be instrumental in the creation of Trismegistos. Together with a database of Demotic papyri by the late Heinz-Josef Thissen, professor of Egyptology at Cologne University, the table of texts collected was at the core of the proposal for the project ‘Multilingualism and Multiculturalism in Graeco-Roman Egypt’ [MaMiGRE], during the course of which Trismegistos would be created (Depauw & Gheldof, 2014).

---

<sup>2</sup> The printed volumes by Leuven scholars all appeared in the series *Studia Hellenistica*, vols. 1-6 (early volumes), vol. 7 (index), vols. 8-9 (addenda et corrigenda). Add also vol. 10 (ethnics) published in 2002. See PP I-X, 1950–2002 in the bibliography.

<sup>3</sup> For the LHPC (now integrated in Trismegistos Collections), see Clarysse & Verreth, 2000; for the Archives project, originally published as the LPHA, now TM Archives, see Clarysse, Vandorpe, & Verreth, 2015; the LDAB is now integrated in TM Texts but also accessible in a separate interface, see LDAB [<http://www.trismegistos.org/ldab>]; the Fayum project is now a part of TM Places, but can still be accessed in its own interface, see Fayum project [<http://www.trismegistos.org/fayum>].

More than just delivering data, however, the PP also inspired the new project in its approach: all-inclusive and interdisciplinary. Even if initially, MaMiGRE was intended to be an Egyptological supplement to the already existing Greek papyrological projects such as the Heidelberger Gesamtverzeichnis griechischer Urkunden aus Ägypten [HGV]<sup>4</sup> and the LDAB, it soon broadened its horizon. Rather than limiting the dataset to papyrology, the Graeco-Roman period and just sources in Egyptian languages and scripts, when TM was launched in 2006 it was meant to be a platform for the study of any type of text dating to the period from 800 BCE to 800 CE, in any language or script and on any writing surface.

In these initial stages, TM Texts still had an important geographical limitation, however, in that it only dealt with Egypt and the Nile Valley. This restriction only disappeared gradually, when after the end of MaMiGRE in 2008 I returned to Leuven and started contemplating the idea of widening our scope to include the entire (Western) ancient world. In 2010, through the mediation of James Cowey, the first contacts were made with the Epigraphische Datenbank Heidelberg [EDH].<sup>5</sup> This eventually allowed us to become a part of the Europeana EAGLE project from 2013 onwards (Orlandi et al., 2017). It also led us to include all Latin inscriptions in TM Texts, a significant increase also in numbers, from roughly 100,000 items (for Egypt), to about 600,000 records. Keeping the interdisciplinary spirit of the PP and TM in mind, however, we also sought cooperation with other projects dealing with the smaller indigenous languages. We thus included 10,000 Etruscan texts through a cooperation with Gerhard Meiser (Meiser, 2014),<sup>6</sup> entered the Messapian (Simone & Marchesini, 2002), Gaulish (*Recueil des Inscriptions Gauloises*, 1985–2002) and Italic (Crawford et al., 2011) evidence on the basis of printed corpora, integrated the Raetic,<sup>7</sup> Ogham (and other Celtic from Britain)<sup>8</sup> and Runic<sup>9</sup> on the basis of existing databases, and also worked together with regional databases such as *Inscriptiones Siciliae* to have exhaustive coverage for specific regions.<sup>10</sup>

TM Texts is still far from complete, however. Our coverage is patchy for languages such as Libyan; Iberian and some other palaeo-Iberian languages are missing completely, as is Punic; we only have the Aramaic material for Egypt, and this is true for most other Semitic languages as well. Our main limitation today, however, is that the Greek inscriptions are still not included, especially for the Greek East

---

4 HGV [<http://aquila.zaw.uni-heidelberg.de>].

5 EDH [<http://edh-www.adw.uni-heidelberg.de>].

6 A re-edition of Rix, 1991.

7 *Thesaurus Inscriptionum Raeticarum* [<http://www.univie.ac.at/raetica/>].

8 Celtic Inscribed Stone Project [<http://www.ucl.ac.uk/archaeology/cisp/>].

9 Runenprojekt Kiel. Sprachwissenschaftliche Datenbank der Runeninschriften im älteren Futhark [<http://www.runenprojekt.uni-kiel.de>].

10 I.Sicily [<http://sicily.classics.ox.ac.uk/>]; see Chapter 19 in this volume. For a full list of our partners, see [[http://www.trismegistos.org/about\\_partners.php](http://www.trismegistos.org/about_partners.php)].

outside Africa. We hope to remedy this in the not too distant future, in cooperation with key research bodies such as the Packard Humanities Institute [PHI]<sup>11</sup> and the Supplementum Epigraphicum Graecum [SEG].<sup>12</sup>

## 15.2 New Techniques & Other Trismegistos Databases

So far the focus has been on the TM Texts database (680,123 records), and rightly so, since the sources lie at the basis of all scholarly research of the history of the ancient world. Nonetheless, Trismegistos also offers other databases, most of which have grown organically from earlier Leuven projects. Trismegistos People is a database of currently 496,702 attestations of people (370,086 records) and personal names (33,325 records) in TM Texts. Although in its current state it cannot really be called a prosopography because people have not been identified systematically across texts (except perhaps for the Ptolemaic period), it clearly builds upon the PP and is currently limited to Egypt. As a systemization of information available in the LDAB, TM Authors deals with ancient authors (5,720 records) and their works (4,847 records – far from complete). At the core of TM Places lies the Fayum project, although it now includes many places (52,130 records) outside Egypt as well, covering both their use as provenance (705,858 records) and their mention in text (217,106 records). The TM Collections database (3,750 records), like its predecessor the LHPC, focuses on the current whereabouts of ancient sources.<sup>13</sup>

Setting up all of these large-scale databases in the last ten years has only been possible because of the availability of full text corpora, new digital processing techniques, and the “exponentiality” permitted by interconnectivity. To start with the former, it was the availability of the full text of Greek papyri in the Duke Databank of Documentary Papyri [DDbDP] that allowed us to develop a Named Entity Recognition [NER] tool to filter out personal names and place names.<sup>14</sup> The NER allowed us to work much faster than would have been possible by purely human input (Depauw & Van Beek, 2009). This is illustrated nicely by the fact that the Demotic evidence, despite the significantly smaller size of the Demotic corpus, is still only partially in the TM People database, whereas the Greek is covered completely – which is entirely due to the fact that Demotic is not available as digital full text. The NER system we set up does not only deal with the typically Greek, and relatively simple, naming system

---

<sup>11</sup> PHI [<http://epigraphy.packhum.org>].

<sup>12</sup> SEG [<http://www.brill.com/publications/online-resources/supplementum-epigraphicum-graecum-online>].

<sup>13</sup> TM People [<http://www.trismegistos.org/ref>]; TM Authors [<http://www.trismegistos.org/authors>]; TM Places [<http://www.trismegistos.org/geo>]; and TM Collections [<http://www.trismegistos.org/coll>].

<sup>14</sup> The DDbDP is now only accessible through the Papyrological Navigator [<http://papyri.info>].



in which most individuals are identified by name and father's name. It can also cope with far more complicated onomastic identifying clusters caused by the Roman *tria nomina* (think of Gaius Iulius Caesar) and the increasingly common addition of mothers, grandfathers etc. to the identification string. Finally, as the DDbDP also included the TM identifiers (discussed further below), we could easily connect the information distilled from the texts to the data that was already available in the TM Texts database: publications, provenance, date, whereabouts etc.

It was this combination of the availability of full text, NER and interconnectivity which allowed (and allows) TM to set up further databases dealing with specific aspects of ancient texts, often in conjunction with other projects and scholars. TM Text Irregularities was developed through a joint effort of Joanne Stolk and myself, to study the corrections both modern editors and ancient authors made in Greek papyri (Depauw & Stolk, 2015).<sup>15</sup> TM Editors sprang from a question to the PAPY mailing-list about papyri edited after 1980, and now identifies over 20,000 modern authors and editors, with special attention to their edition of texts (Depauw & Broux, 2016).<sup>16</sup> TM Abbreviations & Formulae is the result of NER on the full text as available in the Epigraphische Datenbank Clauss-Slaby [EDCS]<sup>17</sup> of Latin inscriptions.<sup>18</sup> It is still under construction, as is the website we are developing on the basis of Ana Blasco's PhD study on the Greek transliteration of Egyptian names (Blasco Torres, 2017). In fact, one could call this last example a double derivate: it builds on the TM People database of names and name variants, which in turn draws in information from TM Texts. Finally, TM Calendar (in cooperation with Sofie Remijsen) is a first attempt at systematizing our date information.<sup>19</sup> We hope to elaborate on this further in the future, in cooperation with projects such as PeriodO and Graph of Dated Objects and Texts [GODOT].<sup>20</sup>

Apart from NER, TM has embraced some other important technical innovations from 2012 onwards. As TM Networks (founded by Yanne Broux) illustrates, we have experimented with what is traditionally called Social Network Analysis [SNA] but now increasingly just network analysis (Broux & Depauw, 2015a).<sup>21</sup> This method of studying connectedness can be used not only to study relations between people, but also places, names or even Demotic epistolary formulae (Broux, 2016; Dogaer & Depauw, 2017). We have also developed a new way of visualising chronological

---

<sup>15</sup> TM Text Irregularities [<http://www.trismegistos.org/textirregularities>].

<sup>16</sup> TM Editors [<http://www.trismegistos.org/edit>].

<sup>17</sup> EDCS [<http://db.edcs.eu>].

<sup>18</sup> TM Abbreviations and Formulae [<http://www.trismegistos.org/abb>].

<sup>19</sup> TM Calendar [<http://www.trismegistos.org/calendar/>].

<sup>20</sup> PeriodO. A Gazetteer of Period Definitions for Linking and Visualizing Data [<http://perio.do>] (see Chapter 16 in this volume); GODOT [<https://godot.date>].

<sup>21</sup> TM Networks [<http://www.trismegistos.org/network>].

evolutions, as it is especially useful to include information from imprecisely dated texts (Van Beek & Depauw, 2013).

One very recent but exciting development has come about through a PhD student, Alek Keersmaekers, whom I co-supervise together with Toon Van Hal (Greek) and Dirk Speelman (corpus linguistics). Starting from the full text of the DDbDP available in GitHub, he has morphologically annotated all the words (part-of-speech tagging and lemmatizing) in XML through a probabilistic model with an accuracy of ca. 95% for non-proper names. Again through a co-operation with TM, he could draw in all the textual metadata, and was also aided by the TM Text Irregularities database for his choice of using the regularized version or the original. We converted his XML to MySQL and made this into the Trismegistos Words database (counting 4,513,494 records) which has become available in January 2018 (Keersmaekers & Depauw, 2018).<sup>22</sup>

### 15.3 The Raison d'Être of Trismegistos

This survey of the roots of the TM project and its development and expansion through new digital techniques may shed some light on the genesis of the project, but I have said preciously little so far about the underlying philosophy of such a broad set of tables or databases.

At the heart of our approach lies the motivation to provide a tool that facilitates access to sources from the ancient world and allows us to study phenomena that transcend disciplinary boundaries. It is only when everything is available in a single system that it is easy to count and quantify. The quantitative method has hitherto been quite marginal in the study of the ancient world, but large corpora of papyri and inscriptions offer interesting new prospects. We have, for instance, revisited the old discussion of the rise of Christianity in the fourth century AD on the basis of the use of Christian names (Depauw & Clarysse, 2013; Depauw & Clarysse, 2015); the increasing use of mother's names in identification clusters (Broux & Depauw, 2015b); the practice of naming your child after a Hellenistic queen (Clarysse & Broux, 2016); or the rise in popularity of double names and hybrid names in the Roman period (Broux, 2015; Dogaer, 2015a; Dogaer, 2015b; Dogaer & Depauw, 2017). In other publications networks, also a form of quantification, are used to study co-occurrence of place names or combinations of epistolary formulae (Broux & Depauw, 2015a). Much more is possible, and I hope that others will start using the data in TM for their own quantitative research.

This brings me to interoperability. From the outset, TM wanted to bring together projects, each collecting data within their scholarly disciplines. TM was never intended to replace projects, if alone for the lack of expertise on most of the languages

---

<sup>22</sup> TM Words [<http://www.trismegistos.org/words>].

and datasets covered by TM Texts. This is also the reason why we, as a rule, do not include the full text itself, nor images of the objects on which the texts are written. Our focus is on (limited) metadata, i.e. information about texts, rather than the texts themselves.

Also, to stimulate cooperation, TM provides stable identifiers for all areas it covers. These identifiers consist of the name of the table or database, and a simple number without meaning that merely identifies the entity and points to information about it in the Trismegistos database. They exist in a human readable format (e.g. TM Nam 1234) or as a “clean” URI (e.g. [<http://www.trismegistos.org/name/1234>]). TM meanwhile has IDs for texts, people, attestations of people, personal names, places, (ancient) authors and their works, (modern) editors, collections, and many more things.

Perhaps the most crucial identifier is the TM Text ID, normally abbreviated as “TM ID” [<http://www.trismegistos.org/text/1234>]. It points to a text or document, in the sense of a set of intentionally related units of linguistically coherent language, written on a physically separate writing surface. The criterion of intentionality is to some extent arbitrary, in the sense that in some cases it is debatable whether two texts actually appear on the same writing surface because their scribes and authors wanted them to. It is, nevertheless, a necessary factor, as otherwise texts appearing on the same object as the result of unrelated reuse would get only a single id. Certainly, in cases where there is no clearly physically separate writing surface (e.g. a desert rock), this would lead to the accumulation of unrelated texts under a single number.

We are very pleased that the Digital Archive for the Study of pre-Islamic Arabian inscriptions [DASI]<sup>23</sup> has agreed to have its material included in Trismegistos. We hope the addition of 7,719 records will make the South Arabian inscriptions better known to scholars of the ancient world, and increase interoperability and standardization. As TM (and other) identifiers spread to as many projects as possible, projects can cooperate and exchange information more easily. In a Linked Open Data Structure, this would permit specialized projects to connect to TM and pull in varied metadata about provenance, date, and publications. This can then be used as background information for the specific topic that forms the focus of attention. In fact, Linked Open Data has the potential to speed up small projects significantly, similar to the development of new tables and databases in TM (Depauw & Dzierzbicka, 2018). Together with other databases such as Pleiades and Pelagios for places or SNAP for people (Simon, Barker, Isaksen, & de Soto Cañamares 2015; Depauw et al., 2017),<sup>24</sup> a graph environment can be created that has great potential to bring knowledge about the ancient world closer to everyone.

---

<sup>23</sup> See Chapter 1 in this volume.

<sup>24</sup> Pleiades [<https://pleiades.stoa.org/>]; Pelagios Commons. Linking the Places of our Past [<http://commons.pelagios.org/>]; Standards for Networking Ancient Prosopographies [<https://snapdrgn.net/>].

## Bibliography

- Blasco Torres, A. (2017). *Representing Foreign Sounds: Greek Transcriptions of Egyptian Anthroponyms from 800 BC to 800 AD* (PhD thesis). Leuven.
- Broux, Y. (2015). *Double Names and Elite Strategy in Roman Egypt* (Studia Hellenistica 54). Leuven: Peeters.
- Broux, Y. (2016). Detecting Settlement Communities in Graeco-Roman Egypt. *Bulletin of the American Society of Papyrologists*, 53, 271–288.
- Broux, Y. & Depauw, M. (2015a). Developing Onomastic Gazetteers and Prosopographies for the Ancient World through Named Entity Recognition and Graph Visualization: Some Examples from Trismegistos People. In L. Aiello & D. McFarland (Eds.), *Social Informatics* (Lecture Notes in Computer Science 8852) (pp. 304–313). Cham: Springer.
- Broux Y. & Depauw, M. (2015b). The Maternal Line in Greek Identification. Signalling Social Status in Roman Egypt (30 BC – AD 400). *Historia. Zeitschrift für Alte Geschichte*, 64, 467–478.
- Clarysse, W. & Broux, Y. (2016). Would You Name Your Child After a Celebrity? Arsinoe, Berenike, Kleopatra, Laodike and Stratonike in the Greco-Roman East. *Zeitschrift für Papyrologie und Epigraphik*, 200, 347–362.
- Clarysse, W. & Verreth, H. (Eds.). (2000). *Papyrus Collections World Wide. 9-10 March 2000 (Brussels - Leuven)*. Brussel: Koninklijke Vlaamse Academie van België voor Wetenschappen en Kunsten.
- Clarysse, W., Vandorpe, K., & Verreth, H. (2015). *Graeco-Roman Archives from the Fayum* (Collectanea Hellenistica 6). Leuven - Paris - Bristol: Peeters.
- Crawford, M., Broadhead, W.M., Clackson, J.P.T., Santangelo, F., Thompson, S., Watmough, M., Bissa, E., & Bodard, G. (2011). *Imagines Italicae: A Corpus of Italic Inscriptions* (Bulletin of the Institute of Classical Studies supplement 110). London: Institute of Classical Studies University of London.
- Depauw, M. & Broux, Y. (2016). Editions and Editors of Greek Papyrological Texts, 1708-2015. *Zeitschrift für Papyrologie und Epigraphik*, 198, 202–210.
- Depauw, M. & Clarysse, W. (2013). How Christian was Fourth Century Egypt? Onomastic Perspectives on Conversion. *Vigiliae Christianae*, 67, 407–435.
- Depauw, M. & Clarysse, W. (2015). Christian Onomastics: A Response to Frankfurter. *Vigiliae Christianae*, 69, 327–329.
- Depauw, M. & Dzierzbicka, D. (2018. In press). Modern Scholars at Work in a Digital and Multidisciplinary Setting. In K. Vandorpe (Ed.), *A Companion to Greco-Roman and Late antique Egypt, 332 BCE – 642 CE* (Blackwell Companions to the Ancient World, forthcoming). Hoboken: Wiley (Blackwell).
- Depauw, M. & Gheldof, T. (2014). Trismegistos. An interdisciplinary Platform for Ancient World Texts and Related Information. In Ł. Bolikowski, V. Casarosa, P. Goodale, N. Houssos, P. Manghi, & J. Schirrwagen (Eds.), *Theory and Practice of Digital Libraries - TPDL 2013 Selected Workshops* (Communications in Computer and Information Science 416) (pp. 40–52). Cham: Springer.
- Depauw, M. & Stolk, J. (2015). Linguistic Variation in Greek Papyri: Towards a New Tool for Quantitative Study. *Greek, Roman, and Byzantine Studies*, 55, 196–220.
- Depauw, M. & Van Beek, B. (2009). People in Greek Documentary Papyri. First Results of a Research Project. *The Journal of Juristic Papyrology*, 39, 31–47.
- Depauw, M., Bodard, G., Cayless, H., Isaksen, L., Lawrence, F., & Rahtz, S. (2017). Standards for Networking Ancient Person-data: Digital Approaches to Problems in Prosopographical Space. *Digital Classics Online*, 3(2), 23–38. Retrieved from [http://journals.ub.uni-heidelberg.de/index.php/dco/article/view/37975], 2017/12/09.

- Dogaer, N. (2015a). Egyptian Names Derived from Foreign Elements: Innovation in Egyptian Onomastic Practice after the Roman Conquest. *Chronique d'Égypte*, 90(180), 360–370. doi: 10.1484/J.CDE.5.110408
- Dogaer, N. (2015b). Greek names with the ending -ιανός/-ianus in Roman Egypt. *The Journal of Juristic Papyrology*, 45, 45–64.
- Dogaer, N. & Depauw, M. (2017). Mapping the Demotic Epistolary Framework through Network Visualisation. *Zeitschrift für Ägyptische Sprache und Altertumskunde*, 144, 173–187.
- Keersmaekers, A. & Depauw, M. (2018). In press. Bringing Together Linguistics and Social History in Automated Text Analysis of Greek Papyri. In A. Novokhatko (Ed.), *Digital Classics III: Re-Thinking Text Analysis* (Classics@, forthcoming). Cambridge, MA: Harvard University.
- Meiser, G. (2014). *Etruskische Texte. Editio minor*. Hamburg: Baar-Verlag.
- Mooren, L. (2001). The automatization of the Prosopographia Ptolemaica. In I. Andorlini, G. Bastianini, M. Manfredi, & G. Menci (Eds.), *Atti del XXII Congresso Internazionale di Papirologia, Firenze, 23-29 agosto 1998* (pp. 995–1008). Firenze: Istituto papirologico G. Vitelli.
- Orlandi, S., Santucci, R., Mambrini, F., & Liuzzo, P. M. (Eds.). (2017). *Digital and Traditional Epigraphy in Context. Proceedings of the EAGLE 2016 International Conference*. Roma: Sapienza Università Editrice. doi: 10.13133/978-88-9377-021-7
- Peremans, W. (1937). *Vreemdelingen en Egyptenaren in Vroeg-Ptolemaeisch Egypte* (Receuil de travaux publiés par les membres des Conférences d'Histoire et de Philologie. 2. Ser., 43). Leuven: Bureaux du recueil - Bibliothèque de l'Université.
- PP I - X (1955–2002). *Prosopographia Ptolemaica* (Studia Hellenistica). Vols. 1–2: Peremans, W. & Van 't Dack, E. (1950 & 1952); vol. 3: Peremans, W., De Meulenaere, H., IJsewijn, J. et al. (1956); vols. 4–5: Peremans, W. & Van 't Dack, E. (1959 & 1963); vol. 6: Peremans, W., Van 't Dack, E., Mooren, L. et al. (1968); vol. 7: Peremans, W., De Meulemeester-Swinne, L., Hauben, H. et al. (1975); vol. 8: Peremans, W., Van 't Dack, E., Mooren, L. et al (1975); vol. 9: Peremans, W., Van 't Dack, E., & Clarysse, W. (1981); vol. 10: La'da, C. (2002).
- Recueil des Inscriptions Gauloises* (1985–2002). Vols. 1 & 2.1: P. Lejeune (1985 & 1988); vol. 2.2: P.-Y. Lambert (2002). Paris: CNRS Éditions.
- Rix, H. (1991). *Etruskische Texte. Editio Minor*. Tübingen: Narr.
- Simon, R., Barker, E., Isaksen, L., & de Soto Cañamares, P. (2015). Linking early geospatial documents, one place at a time: annotation of geographic documents with Recogito. *e-Perimetron*, 10(2), 49–59. Retrieved from [http://oro.open.ac.uk/43613/], 2017/12/09.
- Simone de, C. & Marchesini, S. (2002). *Monumenta linguae Messapicae*. Wiesbaden: Ludwig Reichert.
- Van Beek, B. & Depauw, M. (2013). Quantifying Imprecisely Dated Sources. A New Inclusive Method for Charting Diachronic Change in Graeco-Roman Egypt. *Ancient Society*, 43, 101–114.

Adam Rabinowitz, Ryan Shaw and Patrick Golden

## 16 Making up for Lost Time: Digital Epigraphy, Chronology, and the PeriodO Project

**Abstract:** Digital epigraphy has made great strides toward interoperability and data integration over the last two decades, and Linked Data approaches are now taking advantage of the spatial information associated with inscriptions for new search and visualization tools. The ability to search across epigraphic collections by time, and especially by relative chronologies, lags behind. The PeriodO project has created a Linked Data gazetteer of structured period definitions that facilitates translation between absolute dates and relative chronologies, creating new possibilities for the interoperability of epigraphic collections and their connection with archaeological databases.

**Keywords:** periodization, Linked Open Data, gazetteers, reconciliation, interoperability

### 16.1 The Promise of Digital Epigraphy

The field of epigraphy, with its widely-dispersed body of evidence, its longstanding conventions for description and publication, and its bewildering range of publication venues, has been positioned to benefit from digital approaches since the dawn of the digital age. For the Classical world, this was demonstrated by such early projects as the Packard Humanities Institute digital corpus of Greek inscriptions (Iversen, 2007), and has been confirmed by an array of further efforts spurred on by the rise of the internet. On the most basic level, a digital environment makes it possible to assemble and search across collections of inscriptions that are otherwise scattered in both geographic and bibliographic space. In the last two decades, following the development of the EpiDoc extension of the Text Encoding Initiative to permit the encoding of inscriptions in XML (Cayless et al., 2009; Bodard, 2010), the publication venues themselves have moved online (Reynolds, Roueché, & Bodard, 2007; Bodard, 2008), and the possibilities for the discovery and integration of epigraphic texts have increased exponentially.

At the same time, new digital tools have enhanced the documentation of the physicality of inscriptions, which had long been neglected in publications in

---

**Adam Rabinowitz**, The University of Texas at Austin

**Ryan Shaw, Patrick Golden**, The University of North Carolina at Chapel Hill



© 2018 Adam Rabinowitz, Ryan Shaw and Patrick Golden

This work is licensed under the Creative Commons Attribution 4.0 International license (CC BY 4.0)

favour of textual content. Some of these, like laser scanning, have presented a fairly high technical and financial bar to entry, but others use cheaper and more widely available technologies, such as flatbed scanners or computational photography, to create 2.5D or 3D images (Barmpoutis, Bozia, & Wagman, 2010; Rabinowitz, Schroer, & Mudge, 2010). These representations capture both the text and the materiality of epigraphic objects more fully than photography alone, and with more potential for interactivity. On the other hand, they require more technical investment in online viewing platforms, and their dependence on customized viewers makes them more fragile in the changing online environment. As a result, these techniques have not been incorporated into digital epigraphic practice to the same extent as the EpiDoc text-encoding standard.

Developments in these two areas reflect the traditional consideration of an inscription as a combination of text and object. A third area of digital potential, however, reflects a more recent concern not only with the materiality of inscriptions, but also with their archaeological context. In the 19<sup>th</sup> and early 20<sup>th</sup> centuries, the epigraphic record was valued for the contributions it could make to our understanding of history, and thus the context in which inscriptions were found was usually considered less important than the texts themselves, especially if the stone bearing the inscription had been moved from its original position or reused. More recently, however, the importance of archaeological context for the understanding of inscriptions has been recognized, both within individual sites and on the level of broader regional landscapes (e.g. Holdenried, Roueché, & Scholz, 2014). Fortunately, this recognition has been accompanied not only by an explosion in the online availability of archaeological data, but also in the emergence of a number of projects focused on the aggregation of such data across datasets, like the ARIADNE infrastructure (Niccolucci & Richards, 2013). It is thus increasingly possible to connect aggregations of epigraphic data, such as EAGLE, with aggregations of related archaeological resources, enriching our understanding of the relationship between text and context.

Epigraphic corpora have used space as a central organizational principle since the 19<sup>th</sup> century, from the regional division to the single site. It is therefore not surprising that space and place have offered the easiest point of entry for data integration. Trying to join databases of objects and inscriptions by place-name strings, however, is a futile endeavour: it is not feasible to connect information by strings across databases in a dozen different languages, especially when place-names are often spelled in different ways within a single language. Many projects that seek to create interoperability focus instead on the use of shared external reference points: “gazetteers” that establish the identity of a spatial entity unambiguously, in a standardized and consistently structured format attached to a unique and persistent identifier. By describing metadata values within a database in a semantically transparent fashion, and by including in those descriptions links to persistent identifiers that are themselves described in a semantically transparent fashion, a database manager can plug records in to a wider network of related information.

These are the principles that characterize the Linked Data ecosystem (Heath & Bizer, 2011), and several ancient-world initiatives have already made significant advances by adopting them (Depauw & Gheldof, 2013; Elliott, Heath, & Muccigrosso, 2014; Isaksen et al., 2014). The Pelagios project demonstrates the potential of this approach: its *Recogito* tool associates place-names in texts with entries in gazetteers, while its *Peripleo* browser aggregates data from a variety of datasets that refer to shared historical gazetteers to permit cross-search by ancient place.<sup>1</sup> The datasets aggregated by *Peripleo* already include the *Epigraphische Datenbank Heidelberg*, and more coordinated efforts to integrate inscriptions into the larger Linked Data environment are beginning to materialize (Álvarez, Gómez-Pantoja, & García-Barriocanal, 2011; Blanke et al., 2012). These efforts focus on named entities, which are the most susceptible to disambiguation, unique identification, and manual or automated extraction from text. This work again requires shared points of reference for identification, which are currently provided by spatial gazetteers for place, and are in development for past people (Lawrence & Bodard, 2015; Depauw et al., 2017). Temporal periods, however, despite being the other named entity most frequently encountered by scholars of the past, have until recently been conspicuously absent from this emerging ecosystem.

## 16.2 The Trouble with Time

Both epigraphy and archaeology have long traditions of arranging information according to geographical space, so place-based data aggregation comes very naturally to these disciplines. Both are also deeply engaged in questions of time – but here the two diverge in the nature of their evidence. Inscriptions sit at the intersection between the world of absolute dates, common to textual sources, and the world of relative chronologies based on style, more closely associated with archaeology and art history. On the one hand, calendrical expressions, names and titles of rulers or officials, and particular letter-forms are often very closely dated, to the point where inscriptions, like coins, are used to provide absolute dates for archaeological contexts. On the other hand, inscriptions that lack clearly datable features, or that were produced in periods for which absolute dates are less well-established, are often organized in broader stylistic classes. In some cases, those relative chronologies are the same as those used to classify archaeological material; in others, they were developed specifically for the epigraphic record. In some cases, they are shared widely across multiple geographic regions (for example, the classification “Roman period”); in other cases, they are unique to a single region or language group. And in some cases, these relative chronologies are attached to absolute dates, while in

---

<sup>1</sup> [<http://recogito.pelagios.org/>]; [<http://peripleo.pelagios.org/>].



others their dating is either left open or inferred from absolute dates ascribed to the inscriptions themselves.

Archaeological remains, on the other hand, are much more commonly classified by relative chronologies based on a complicated and idiosyncratic combination of historical, stylistic, and material features. The defining characteristic of these chronologies is their division into “periods”, blocks of time that the scholarly community assumes to be characterized by distinct and consistent qualities or phenomena. While these periods can appear to be fairly consistent across regions and projects – “Roman”, for example, seems like a transparent term at first glance – the apparent agreement masks a vast number of chronological inconsistencies and disagreements based on factors like geography (“Roman” in the UK does not have the same temporal range as “Roman” in Italy, for example) or school of thought (where does “Roman” stop and “Late Antiquity” begin?).

As a result, although archaeologists, epigraphers, and historians alike group material by time as often as they do by space, time has resisted the integration strategies applied so effectively to space by the Pelagios project. Variation in the usage and meaning of period terms makes it difficult to integrate archaeological records chronologically across multiple databases, and it makes it even more difficult to integrate those records with the contents of epigraphic databases, which often eschew periodization altogether, or use it only in the absence of tight absolute chronologies. Some epigraphers might not see this as a real problem: after all, absolute dates can be easily searched both within and across databases, as long as some basic standards for date format are observed. But to ignore the issue is to discard one of the greatest benefits of the emerging digital ecosystem for the study of the past: the combination of different strands of evidence to create a new understanding of ancient societies. In some cases, integration might even lead us to reconsider long-standing knowledge categories. What would we find, for example, if we could compare current epigraphic work to redefine the meaning of “Late Antiquity” (Tantillo, 2017) with objects described with terms analogous to “Late Antique” across multiple languages and databases? Furthermore, better strategies for navigating between relative and absolute dating systems might help to expand context for inscriptions currently isolated within idiosyncratic local chronologies.

### 16.3 The PeriodO Temporal Gazetteer

The reconciliation of relative and absolute chronologies, and the clarification of scholarly usage of period terms, is the goal of the PeriodO project.<sup>2</sup> PeriodO offers a Linked Data gazetteer, not of spatial entities, but of definitions of periods located in both space and time. It emerged from the recognition that the spatial and temporal coordinates of period terms, as these terms are used in the study of the human past, are deeply entangled, and that the terms themselves are discursive constructs subject to disagreement and diachronic change (Morris, 1997; Rabinowitz, 2014; Rabinowitz et al., 2016; Kotsonas, 2016). On a chronological (and, arguably, phenomenological) level, there is no single “Roman period” in modern scholarship or datasets: there are a series of related “Roman” periods with different temporal boundaries in different places, and if we want to be able to aggregate data along a temporal axis, it is critical for scholars or data-managers to be able to make transparent statements about which of those meanings of “Roman period” is in play in a particular context. The PeriodO project considers three pieces of information to be critical for a transparent period definition: coordinates – even vague coordinates – in time (an earliest start and a latest stop); coordinates in space (in what part of the world the term is applied with that chronological meaning); and an authoritative source for the association of those coordinates with that period term (Figure 16.1). By modelling both sources and definitions as structured data, and by providing both with unique, persistent identifiers, PeriodO makes it possible for a dataset to make an unambiguous statement about its usage of a given period term (“By ‘Archaic’, we mean the period between 700 BC and 480 BC within the bounds of modern Greece and Turkey, as defined by scholar X”). This in turn makes it easier to visualize and search the contents of that dataset by both time and space, and to understand how the chronology used in one dataset relates to the chronology in another, which might assign different dates to “Archaic” or use a different term (e.g. “Orientalizing”) for part or all of the same date range.

In documenting usage through the collection and modelling of period definitions, PeriodO does not intend to create a centralized, authoritative, prescriptive vocabulary for periods. Instead, the set of required attributes are meant to encourage multivocality: as long as a definition has a date range, a spatial extent, and an authority, and as long as it is not identical to an existing definition in the dataset, it can be added on an equal footing with other definitions. Although the initial content of the dataset was gathered by the project team from published work and from the formal vocabularies contributed by a group of generous partners, our goal is to expand that content in the future through user submissions. If a user interested in deploying PeriodO period

---

<sup>2</sup> [<http://perio.do>]; the permalink for the client interface is [<http://n2t.net/ark:/99152/p0>]. PeriodO has been generously funded by grants from the US National Endowment for the Humanities (grant HD-51864-14) and the US Institute of Museum and Library Services (grant LG-70-16-0009-16).

identifiers in a dataset does not find definitions that match his or her own, new definitions can be submitted to the dataset as a data “patch” that is then merged into the “canonical” dataset (Shaw et al., 2016). As the user community grows, and as the project team continues to add periodizations from new disciplines and more diverse sources (including works written in the 18<sup>th</sup> or 19<sup>th</sup> centuries), we hope that the dataset will serve not just as a source of structured temporal data and identifiers, but also as a representation of the broader scholarly discourse about periodization. In order to make this possible, the dataset includes not only sources and definitions, but also formal links between definitions (specifically, that a period definition is broader or narrower than another definition from the same source, or that a definition in one source is derived from a definition in another source, or that a definition is the same as that described by a Linked Data identifier in another dataset), and a full provenance history describing who submitted data to the dataset, who approved it for inclusion, and when it was merged (Golden & Shaw, 2016).

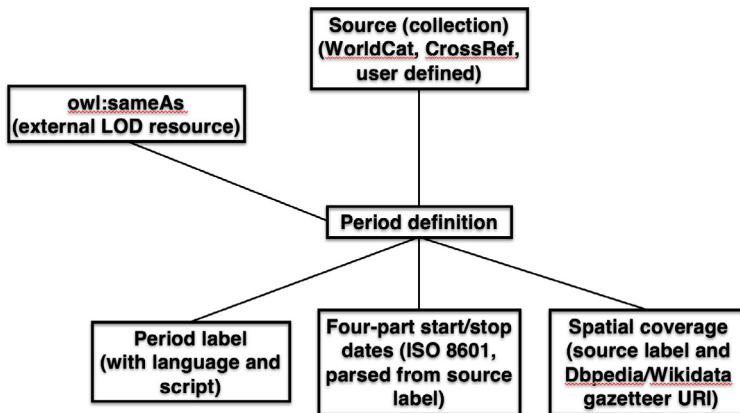


Figure 16.1: Diagram of PeriodO data model

### 16.3.1 PeriodO and Digital Epigraphy

These features – the embrace of multiple definitions of periods, extensibility by a user community, attention to scholarly provenance and intellectual genealogy – make PeriodO particularly useful for the integration of digital epigraphic collections into a Linked Data ecosystem. By avoiding a centralized vocabulary, it allows the discipline to document the different period definitions used by epigraphers across dozens of countries over the last several hundred years, while facilitating the reconciliation of a wide range of locally- or regionally-specific periodizations used in current databases. The ability to match periodized material in one database with periodized material

in another, by period term or date-range or both, offers significant advantages for scholarship both within the field of epigraphy and outside it.

Within the field, for both experienced and novice epigraphers, there are times when it is useful to assemble a set of inscriptions that are contemporary in date across several corpora. This has always been difficult with the printed record. Although inscriptions in a specific publication or fascicule are usually arranged in chronological order, they are also published as they come to light, which means that in the best case, inscriptions of the same general period can be spread across several different volumes in a single series (more frequently, they are spread across multiple series and specialised venues). The situation is somewhat eased in digital collections, in which records can be reorganized according to any criteria included in metadata and considered by the database designer. But the way in which dating criteria are considered differs widely from collection to collection. The Epigraphische Datenbank Heidelberg, for example, has long allowed search by periods derived from Roman political history, but until recently it simply used those periods as a proxy for absolute date ranges.<sup>3</sup> By contrast, the Europeana EAGLE database, which aggregates inscriptions from several different epigraphic collections, does include “period” as a metadata attribute, but does not have a period search facet.<sup>4</sup> EAGLE and the online publication of the Aphrodisias inscriptions<sup>5</sup> both allow searching by absolute dates, as a date range alone (for the former) or by either date range or century (for the latter). The PHI database of Greek epigraphy includes absolute dating information drawn from the published corpora, but does not allow any searching or browsing by date or date range.<sup>6</sup> Other online collections include periods as metadata attributes, but because of uncertainty about the relation between relative and absolute chronologies do not include any date information.

A metadata attribute that points to an identifier in an external gazetteer for a structured spatiotemporal representation of a period term has the potential to bring some order to this chaos. This is especially true when that identifier also offers a transparent record not only of authority, but also of uncertainty. The date ranges associated with PeriodO definitions are parsed from date expressions in the original

---

<sup>3</sup> [<http://edh-www.adw.uni-heidelberg.de/>]; see search interface at [<http://edh-www.adw.uni-heidelberg.de/inschrift/suche>], where a search by “Historische Periode” “entspricht einer Datierungssuche mit durch Jahreszahlen definierten Zeiträumen”. In December 2017, however, the EDH added period identifiers from PeriodO to the metadata for its dated inscriptions.

<sup>4</sup> [<https://www.eagle-network.eu/>] (Liuzzo, 2014, with metadata specification at [https://www.eagle-network.eu/wp-content/uploads/2013/06/EAGLE\\_D3.1\\_EAGLE-metadata-model-specification\\_v1.1.pdf](https://www.eagle-network.eu/wp-content/uploads/2013/06/EAGLE_D3.1_EAGLE-metadata-model-specification_v1.1.pdf)); metadata vocabularies corresponding to the notion of historical period are divided into “reign of emperors” [<https://www.eagle-network.eu/voc/dates/lod/22.html>] and more general “periods” [<https://www.eagle-network.eu/voc/dates/lod/8.html>].

<sup>5</sup> [<http://insaph.kcl.ac.uk/iaph2007/index.html>].

<sup>6</sup> [<http://epigraphy.packhum.org/>].

source, but we also retain the original date labels, which can be as vague as “around the middle of the second century BC”. Furthermore, the proleptic Gregorian calendar dates, expressed according to the ISO8601 standard and the OWL-Time ontology,<sup>7</sup> can be structured as a four-part date range, with earliest/latest start and earliest/latest stop, in order to preserve fuzzy chronological boundaries while still allowing date-based search. PeriodO identifiers thus make it easier to search within a single dataset by both date range and period term, while facilitating cross-searching and aggregation across different datasets that share the gazetteer as a common reference point.

Perhaps even more importantly, reference to a shared temporal gazetteer provides a bridge between inscriptions with absolute dates and archaeological material classified by period, enabling union searches that return both kinds of records. The Pelagios Project’s Peripleo browser already provides a model for such searches, but since it is only beginning to incorporate periods as a search facet, the current timeline filter is useful primarily for objects with absolute dates, like coins. With the addition of shared external reference points for structured-data representations of periods, we will move closer to a fully integrated spatiotemporal search, within which a single bounded query could return Palmyrene sculptures contemporary with Palmyrene epigraphy, or Pompeiian graffiti together with Flavian-period wall-painting. Such combinations of the material context and the epigraphic record have the potential to shed new light on both sides.

### 16.3.2 Using the PeriodO Gazetteer in Epigraphic Corpora

Before we can reach this point, however, there are more mundane considerations. The most pressing involves the sea of data a user must navigate in PeriodO, which now contains more than 5,000 definitions, many of them referring to the same or similar concepts. The PeriodO project provides user documentation both on its current homepage and in a Github repository.<sup>8</sup> While the project’s online documentation should be seen as the definitive guide, it is nevertheless useful to discuss the structure of the dataset and how PeriodO URIs can be added to epigraphic collections.

#### 16.3.2.1 Technical Specifications

The PeriodO dataset is, at the core, a single plain-text file in the JavaScript Object Notation (JSON) format, interpretable as RDF via the JSON-LD (JSON for Linking Data) standard. The dataset is described using terms from standardized vocabularies

---

<sup>7</sup> [<https://www.w3.org/TR/owl-time/>].

<sup>8</sup> [<https://github.com/periodo>].

including the Simple Knowledge Organization System (SKOS), the Time Ontology in OWL (OWL-Time), and the Dublin Core Metadata Terms. The dataset as a whole is identified by an Archival Resource Key (ARK) identifier from the California Digital Library EZID system. Persistent HTTP resource identifiers (URIs) for each period collection (the authoritative source for one or more period definitions) and each period definition are provided via the EZID Name-to-Thing (N2T) resolver, which works with the ARK ID system (Kunze & Rodgers, 2013).

Acronyms and jargon aside, this means that the dataset is lightweight, hierarchical in structure, standard in format, human- and machine-readable, and provided with persistent, globally unique identifiers for its contents. Snapshots of the dataset will be preserved in a long-term institutional repository under open-access terms, so that if the web front-end ever ceases to work, the ARK ID will always point to a final version of the dataset, and the identifiers will always remain globally unique and persistent, even if the URI cannot be resolved as a URL. The structure of the PeriodO dataset also means that it is easy to download and reuse, adapt, and repurpose it, or to run it from a local server. Long-term preservation will be handled by the University of Texas Libraries, so there is very little risk that access to PeriodO data will be compromised in the foreseeable future.

### 16.3.2.2 Reconciliation

While it is possible to find period definitions by browsing the dataset through the PeriodO client, and to add their URIs to an epigraphic dataset manually by copy-pasting, it is not the most efficient process when a large number of period terms are involved. A reconciliation service is a digital tool that uses an algorithm to automatically match values in one dataset (for example, a column containing place-names in a spreadsheet) to similar values in another (for example, a gazetteer of historical places).<sup>9</sup> Such services can be web-based, like the Geocollider tool recently developed to facilitate the matching of place-names in user-submitted structured data with Pleiades identifiers,<sup>10</sup> or they can be integrated into another data-cleaning tool like OpenRefine.<sup>11</sup> This makes it easier for a data manager to match a large number of values at once to an external reference point, rather than copying and pasting one URI at a time. PeriodO has developed a reconciliation service for OpenRefine, instructions for which are available on Github.<sup>12</sup> Using the PeriodO reconciler, a user can match period terms in a structured-data document (in formats such as CSV, XML, JSON, etc.) to period definitions in PeriodO, using not only the term itself but also values in other

---

<sup>9</sup> [<https://github.com/OpenRefine/OpenRefine/wiki/Reconciliation>].

<sup>10</sup> [<http://geocollider-sinatra.herokuapp.com/>].

<sup>11</sup> [<http://openrefine.org/>]. The Geocollider tool is also offered as a service through OpenRefine.

<sup>12</sup> [<https://github.com/periodo/periodo-reconciler>].

columns like start or stop date and spatial coverage to refine the matching process (Figure 16.2). If, then, the manager of an epigraphic database wishes to add PeriodO identifiers to periodized records or a list of period terms, the reconciler makes the process simpler and faster.

**Figure 16.2:** Using the PeriodO reconciler with OpenRefine to match period terms from the EDH search page to period definitions in the gazetteer

### 16.3.2.3 Adding Data to the Gazetteer

Inevitably, however, some of the periods used in any given dataset will not match any existing values in PeriodO. A near-match might be sufficient for a data manager in some cases, but in others there may be a local period definition that has to be expressed as-is. The PeriodO gazetteer has been designed with the expectation that new users will find new gaps, and therefore it has a process to allow users to fill in missing pieces. The web interface for the platform allows the user to create and edit local period databases, either using collections synced from the “canonical” dataset or generating entirely new collections and definitions. Any user with an ORCID<sup>13</sup> may use it to log in to the PeriodO client and submit one of these local databases with new or revised period entries as a patch to the server. If new definitions meet the basic requirements of the dataset (authority and spatiotemporal coordinates), and if they are formatted correctly (specifically, if they include the original wording and values

<sup>13</sup> [<https://orcid.org/>].

used by the source cited for spatial coverage and dates, rather than an interpretation of either by the user), the patch is merged with the “canonical” dataset on the PeriodO server, and persistent URIs are minted for the new definitions. The patch process not only guarantees that new data will meet the criteria and formatting expectations of the platform, but provides a clear documentation trail for the process of submission and approval. This trail itself, including the actors involved, is modelled using the Provenance Ontology and added to the PeriodO dataset, so that any definition can be associated with the individuals who proposed or approved it.

#### 16.3.2.4 EpiDoc Guidelines

The previous paragraphs have described how the manager of a digital epigraphic collection can associate PeriodO URIs with local period terms contained in a spreadsheet or XML document. For collections that are already being expressed in the EpiDoc extension to TEI-XML, it is also important to understand how PeriodO URIs should be represented in that convention. Fortunately, the EpiDoc extension has a property class for named historical periods, which is described in the current version of the EpiDoc guidelines.<sup>14</sup> Such periods can be encoded in an EpiDoc representation within the “origDate” element using the “period” attribute, according to the example given:

```
<origDate notBefore="-0332"
notAfter="-0200" precision="medium"
period="http://n2t.net/ark:/99152/p0m63njc4hd"    evidence="lettering">    Early
Hellenistic (lettering)</origDate>
```

PeriodO is accepted in the convention as an authoritative source of URIs for period terms in this context.<sup>15</sup>

## 16.4 Conclusions

Just as the shift from print to digital epigraphic corpora opened a world of new possibilities for searching and aggregation in the 1980s and early 1990s, and just as the shift from CD-ROMs to online databases did this again for the discipline in the early 2000s, the maturation of semantic-web approaches in recent years has begun to reveal the potential of Linked Data for discovery and data integration. This is an exciting development, since it promises to allow us to find unexpected

---

<sup>14</sup> [<http://www.stoa.org/epidoc/gl/latest/>].

<sup>15</sup> [<http://www.stoa.org/epidoc/gl/latest/supp-historigdate.html>].



conjunctions between inscriptions in different collections, and between inscriptions and archaeological material, in ways that were barely imaginable a few decades ago. With the Pelagios project, the spatial component of this process of linking and aggregation has taken off. The temporal component still lags behind, however, simply because – unlike places, which exist in physical space – periods are discursive constructs that emerge from the needs of scholarly studies of the past to create order. As discursive constructs, they change over time and inspire revision, disagreement, and critique. This makes them difficult to manage in a structured-data environment: capturing the diversity of usage can create an impression of chaos, while smoothing out disagreement both excludes critique and erases some of the history of historical disciplines. One can see why absolute dates or generic period expressions might be more attractive for managers of digital epigraphic collections.

We hope, however, that we have shown some of the benefits that come with entry into the fray, and the goal of the PeriodO project is to continue to make it easier to do so. If the digital epigraphic community begins to include periods systematically in its data structures, it will be rewarded with better interoperability across datasets, better ways to find information about inscriptions, and – perhaps most importantly of all – better opportunities to reunite inscribed texts with archaeological context at various scales. The flexibility of the PeriodO gazetteer should be able to meet the needs of a wide range of period uses in epigraphic corpora, from the relatively straightforward chronology of the Inscriptions of Israel/Palestine, which is largely satisfied with the period definitions used by the Levantine Ceramics Project, to the highly specific linguistic/stylistic periods that appear in some of the corpora of the Digital Archive for the Study of pre-Islamic Arabian Inscriptions.<sup>16</sup> While the usefulness of period metadata may not appear immediately to the early adopters, it will become increasingly evident as more collections incorporate it and as temporal search and visualization tools become more robust. Today we cannot imagine how we managed without the PHI database of Greek epigraphy or the EpiDoc standard; tomorrow, we will not remember what it was like to be able to search easily across dozens of epigraphic collections for Archaic inscriptions alone, or visualize on a map and timeline how different corpora differ in their definitions of “Late Antiquity”. The transparent association of period definitions with material with absolute dates, like inscriptions, may even lead us to a fundamental reconsideration of the way we periodize the past.

---

<sup>16</sup> [<http://cds.library.brown.edu/projects/Inscriptions/>]; [<http://dasi.cnr.it/>].

## Bibliography

- Álvarez, F.-L., Gómez-Pantoja, J.-L., & García-Barriocanal, E. (2011). From Relational Databases to Linked Data in Epigraphy: Hispania Epigraphica Online. In E. García-Barriocanal, Z. Cebeci, M. Okur, & A. Öztürk (Eds.), *Metadata and Semantic Research 5th International Conference, MTSR 2011, Izmir, Turkey, October 12-14, 2011. Proceedings* (pp. 225–233). Berlin-Heidelberg: Springer.
- Bampoutis, A., Bozia, E., & Wagman, R.S. (2010). A novel framework for 3D reconstruction and analysis of ancient inscriptions. *Machine Vision and Applications*, 21(6), 989–998. doi: 10.1007/s00138-009-0198-7
- Blanke, T., Bodard, G., Bryant, M., Dunn, S., Hedges, M., Jackson, M., & Scott, D. (2012). Linked data for humanities research – The SPQR experiment. In *2012 6th IEEE International Conference on Digital Ecosystems and Technologies (DEST) Complex Environment Engineering* (pp. 1–6). doi: 10.1109/DEST.2012.6227932
- Bodard, G. (2008). The Inscriptions of Aphrodisias as electronic publication: A user's perspective and a proposed paradigm. *Digital Medievalist*, 4. doi: 10.16995/dm.19
- Bodard, G. (2010). EpiDoc: Epigraphic documents in XML for publication and interchange. In F. Feraudi-Gruénais (Ed.), *Latin on Stone: Epigraphic Research and Electronic Archives* (pp. 101–118). Lanham, MD: Lexington Books.
- Cayless, H., Roueché, C., Elliott, T., & Bodard, G. (2009). Epigraphy in 2017. *Digital Humanities Quarterly*, 3(1). Retrieved from [http://www.digitalhumanities.org/dhq/vol/3/1/000030/000030.html], 2017/12/10.
- Depauw, M., Bodard, G., Cayless, H., Isaksen, L., Lawrence, F., & Rahtz, S. (2017). Standards for Networking Ancient Person data: Digital approaches to problems in prosopographical space. *Digital Classics Online*, 3(2), 28–43. doi: 10.11588/dco.2017.0.37975
- Depauw, M. & Gheldof, T. (2013). Trismegistos: An Interdisciplinary Platform for Ancient World Texts and Related Information. In *Theory and Practice of Digital Libraries – TPD 2013 Selected Workshops* (pp. 40–52). Cham: Springer. doi: 10.1007/978-3-319-08425-1\_5
- Elliott, T., Heath, S., & Muccigrosso, J. (2014). *Current Practice in Linked Open Data for the Ancient World, ISAW Papers*, 7. New York: Institute for the Study of the Ancient World, New York University; Princeton University Press. Retrieved from [http://dlib.nyu.edu/awdl/isaw/isaw-papers/7/], 2017/12/10.
- Golden, P. & Shaw, R. (2016). Nanopublication beyond the sciences: the PeriodO period gazetteer. *PeerJ Computer Science*, 2(e44). doi: 10.7717/peerj-cs.44
- Heath, T. & Bizer, C. (2011). *Linked Data: Evolving the Web into a Global Data Space* (1st ed., Vol. 1). Morgan & Claypool.
- Holdenried, M., Roueché, C., & Scholz, M. (2014). Digital epigraphy in its archaeological context: the case of Metropolis, Magnesia, and Apollonia. In B. Dreyer (Ed.), *Die Surveys im Hermos- und Kaystrostal und die Grabungen an den Thermen von Metropolis (Ionien) sowie am Stadion von Magnesia am Mäander* (pp. 163–186). Muenster: LIT Verlag Münster.
- Isaksen, L., Simon, R., Barker, E.T.E., & de Soto Cañamares, P. (2014). Pelagios and the emerging graph of ancient world data. In *Web Sci '14. Proceedings of the 2014 ACM conference on Web science* (pp. 197–201). ACM.
- Iversen, P.A. (2007). The Packard Humanities Institute (PHI) Greek Epigraphy Project and the Revolution in Greek Epigraphy. *Abgadiyat*, 2(1), 51–55. doi: 10.1163/2213860907X00057
- Kotsonas, A. (2016). Politics of periodization and the archaeology of early Greece. *American Journal of Archaeology*, 120(2), 239–270.
- Kunze, J. & Rodgers, R. (2013). The ARK Identifier Scheme. Retrieved from [https://tools.ietf.org/html/draft-kunze-ark-18], 2017/12/10.

- Lawrence, K.F. & Bodard, G. (2015). Prosopography is Greek for Facebook: The SNAP:DRGN Project. In *WebSci '15. Proceedings of the ACM Web Science Conference* (p. 44:1–44:2). New York, NY, USA: ACM. doi: 10.1145/2786451.2786496
- Liuzzo, P.M. (2014). The Europeana Network of Ancient Greek and Latin epigraphy (EAGLE). *ISAW Papers*, 7(12). Retrieved from [<http://dlib.nyu.edu/awdl/isaw/isaw-papers/7/liuzzo/>], 2017/12/10.
- Morris, I. (1997). Periodization and the heroes: inventing a Dark Age. In M. Golden & P. Toohey (Eds.), *Inventing Ancient Culture. Historicism, Periodization, and the Ancient World* (pp. 96–131). London and New York: Routledge.
- Niccolucci, F. & Richards, J. (2013). ARIADNE: Advanced research infrastructures for archaeological dataset networking in Europe. *International Journal of Humanities and Arts Computing*, 7(1–2), 70–88. doi: 10.3366/ijhac.2013.0082
- Rabinowitz, A. (2014). It's about time: historical periodization and Linked Ancient World Data. *ISAW Papers*, 7(22). Retrieved from [<http://dlib.nyu.edu/awdl/isaw/isaw-papers/7/rabinowitz/>], 2017/12/10.
- Rabinowitz, A., Schroer, C., & Mudge, M. (2010). Grass-roots imaging: a case study in sustainable heritage documentation at Chersonesos, Ukraine. In B. Frischer & L. Fisher (Eds.), *Making History Interactive: Proceedings of CAA 2009*. Budapest: Archaeolingua.
- Rabinowitz, A., Shaw, R., Buchanan, S., Golden, P., & Kansa, E. (2016). Making sense of the ways we make sense of the past: The PeriodO project. *Bulletin of the Institute of Classical Studies*, 59(2), 42–55. doi: 10.1111/j.2041-5370.2016.12037.x
- Reynolds, J., Roueché, C., & Bodard, G. (2007). *Inscriptions of Aphrodisias*. Retrieved from [<http://insaph.kcl.ac.uk/iaph2007>], 2017/12/10.
- Shaw, R., Rabinowitz, A., Golden, P., & Kansa, E. (2016). A sharing-oriented design strategy for networked knowledge organization systems. *International Journal on Digital Libraries*, 17(1), 49–61. doi: 10.1007/s00799-015-0164-0
- Tantillo, I. (2017). Defining Late Antiquity through Epigraphy? In R.L. Testa (Ed.), *Late Antiquity in Contemporary Debate* (pp. 56–79). Newcastle upon Tyne: Cambridge Scholars Publishing.

Pietro M. Liuzzo

## 17 EAGLE Continued: IDEA. The International Digital Epigraphy Association

**Abstract:** Few disciplines can boast of having digitized almost the entirety of the documents they are interested in, and to have so many scholars active in digitization projects, as in Greek and Latin epigraphy (Orlandi et al., 2014; Orlandi et al., 2017). This paper will present some of the methodological issues faced by the Europeana network for Ancient Greek and Latin Epigraphy, before and after the end of the project when its activities were moved to the International Digital Epigraphy Association. It will give some examples to demonstrate how the above-mentioned achievement is far from being enough to support real user cases. Particularly, problems of mapping will be presented with an evaluation of the current quality of the data, and some hints to the continuing work of the IDEA association for the EAGLE portal and associated resources.

**Keywords:** Greek epigraphy, Latin epigraphy, up-conversion, collaboration, Epigraphy.info

### 17.1 The EAGLE Project Steps

#### 17.1.1 The EAGLE Aggregator

Within the EAGLE project (Europeana Network for Ancient Greek and Latin Epigraphy) a model was established, based on the principles of the TEI/EpiDoc standard (Amato et al., 2013; Manghi et al., 2015) that was able to guarantee easy mapping to the CIDOC-CRM<sup>1</sup> and to EDM (Europeana Data Model) for harvesting purposes.<sup>2</sup> As a result, this work has made possible not only the development of the EAGLE portal, with its search functionalities across data from several different sources, but has also allowed the

---

1 [<http://www.cidoc-crm.org/>].

2 [<https://pro.europeana.eu/resources/standardization-tools/edm-documentation>].

---

Pietro M. Liuzzo, Universität Hamburg



© 2018 Pietro M. Liuzzo

This work is licensed under the Creative Commons Attribution 4.0 International license (CC BY 4.0)

creation of a large network of partners, and produced additional resources such as the EAGLE Vocabularies and the Virtual Exhibition.<sup>3</sup>

### 17.1.2 The EAGLE Portal

The EAGLE portal is an entry point to the content of the collections participating in the project, not only the databases of the original EAGLE – Electronic Archive of Greek and Latin Epigraphy (Epigraphic Database Heidelberg, Epigraphic Database Rome, Epigraphic Database Bari and Hispania Epigraphica Online),<sup>4</sup> but also many other projects like Ubi erat lupa, Last Statues of Antiquity, the Ancient Graffiti Project and, at the end, part of the Epigraphic Database Clausus Slaby.<sup>5</sup> It can be confidently stated that almost all of the existing digital epigraphic projects were in some way linked with the EAGLE project during and after its lifetime. Here are some features of the EAGLE portal in brief:

1. it has a data aggregator that mainly harvests EpiDoc XML exports of a rich, but minimal, set of information from the contributing partners. It also makes them searchable in one place for end users, as well as harvestable for Europeana, the European portal of cultural heritage.<sup>6</sup> Anyone can make its contents part of EAGLE. Anyone can reuse this data,<sup>7</sup>
2. it groups and organizes results based on a unique ID given by Trismegistos (TM), thus offering parallel results for one text,
3. thanks to an image based search system, it makes the collections searchable, for the first time, also by images,
4. it has harmonized vocabularies in use for several descriptive fields,<sup>8</sup>
5. it has developed – and is linked to – a set of services, like a big collection of translations of inscriptions (Bigi, 2014),<sup>9</sup> a storytelling application<sup>10</sup> and a Virtual Exhibition, “Signs of Life” which collects images, 3D models, infographics and many other types of materials to give an introduction to non-experts (Liuzzo, Mambrini, & Franck, 2017).

---

<sup>3</sup> An early version of this paper was presented with Silvia Orlandi at the international congress of Latin and Greek epigraphy in Wien, 2017/8/29.

<sup>4</sup> [[http://www.eagle-eagle.it/Italiano/index\\_it.htm](http://www.eagle-eagle.it/Italiano/index_it.htm)].

<sup>5</sup> [<https://www.eagle-network.eu/eagle-project/partners/>].

<sup>6</sup> [<https://www.europeana.eu>].

<sup>7</sup> Although this is never as easy as one would like, and requires some work from the system maintainers. The IDEA association carries out this work as part of its mission for member institutions. See below for an example.

<sup>8</sup> [<https://www.eagle-network.eu/resources/vocabularies/>].

<sup>9</sup> The Eagle Media Wiki for Translations of inscriptions [[https://wiki.eagle-network.eu/wiki/Main\\_Page](https://wiki.eagle-network.eu/wiki/Main_Page)].

<sup>10</sup> [<https://www.eagle-network.eu/resources/flagship-storytelling-app/>].

Both the portal and the services created by EAGLE are still working after the official end of the European project in March 2016. However, they need to be maintained, constantly updated and possibly improved, as the EAGLE experience has also highlighted the limits and the problems of digital resources.

## 17.2 IDEA

After the end of the EAGLE project, IDEA, the International Digital Epigraphy Association, was founded, with the aim of maintaining the EAGLE resources and continuing on the path of cooperation and integration of resources. It also aimed to cross the boundaries of single projects and move towards the creation of a Epigraphy.info resource (Feraudi-Gruénais & Grieshaber, 2016), based on the model already used by papyrologists.<sup>11</sup> IDEA has, as its primary aim, to continue the networking efforts of the EAGLE project and to maintain its outputs, with a very practical approach: keep the EAGLE portal infrastructure running, together with its functionalities, supporting members who want to contribute, advising new projects on what is and isn't available, keeping an eye on the developments in the field and sharing this knowledge to increase the possibility of more effective and organized work on digital epigraphy.

IDEA currently supports its members and prospective members in a range of activities, from data curation and consultancy on how to set up new digital epigraphic projects, to the upload of new data from existing content providers who continue to update their resources locally.

Actual activity for the current year included, for example:

- aggregation of data from existing partners still actively updating their resources (occasional and not systematic or planned, due to lack of resources),
- continued collaboration with network members and with active projects (e.g. Pondera project,<sup>12</sup> IG Cyr and GVCyr,<sup>13</sup> Iscrizioni Latine Arcaiche<sup>14</sup>),
- server migration and maintenance,
- updates to the EAGLE vocabularies.<sup>15</sup>

---

<sup>11</sup> The Papyrological Navigator and Editor [<http://papyri.info/>]. This model has been superseded by a more distributed data exchange model since the Epigraphy.info meeting held in Heidelberg 21–23 March 2018.

<sup>12</sup> [<https://pondera.incal.ucl.ac.be/>].

<sup>13</sup> [<https://igcyr.unibo.it/>].

<sup>14</sup> This project is not yet online.

<sup>15</sup> [<https://www.eagle-network.eu/resources/vocabularies/>].

### 17.3 Methodological Issues Faced During EAGLE

Mapping and harmonizing the data (as far as possible) from the databases to a minimum set of standardized information, was the primary aim of one of the working groups in EAGLE (Liuzzo, 2017). EpiDoc was the obvious choice, but because of the scope and obligations of the project there was no conversion of the existing databases to an XML workflow. Rather, an additional workflow was generated to export this format for the purposes of aggregation and to allow a common portal to search across databases.

This process meant that each participating database or project had to contribute an export of its data, produced with its resources or those common to the project, validating to the EAGLE schema. This was a stricter version of the EpiDoc schema from which it was generated, with a minimal set of information required by the common definition.

The efforts here went into making this limited information, packed into a strict schema originally intended for database and aggregation purposes, as rich as possible. This would allow further reuse, demonstrating its usefulness as a large corpus of disambiguated information.

We focused on the alignment and harmonization of the vocabularies used for the descriptions of inscriptions and on the up-conversion of the string text into XML (Liuzzo, Fasolini, & Rocco, 2014).

The first task began with the acquisition of all lists used by the partners. We attributed IDs to each concept and then aligned the terms used, marking the language in which they appeared. We immediately faced decisions, such as that of the “main language”. The vocabularies still claim to be in English, although it was expressly declared that the choice of the language for the main label for each concept would not be in one of the many languages represented in the network, but rather prioritize those terms that also had an associated definition. Thus, in the tabular view of the full vocabulary one can see terms in Latin, English and German as the main label (Figure 17.1).

Programmatically speaking, this is something of a problem. However, it better reflects the reality, where a translation of a concept from Italian to English does not correspond to how an English-speaking project labels that concept. There are many reasons for this, such as, the different definition in the context taken into consideration of a certain text typology, or the major or minor degree of precision in the labelling of types. The EAGLE vocabularies want to be inclusive, rather than selective, and help the alignment and connection of entities rather than forcing a denomination or a language to any description. These vocabularies continued to raise interest and continued to receive contributions after the end of the project, especially from the Ancient Graffiti project<sup>16</sup> and the I.Sicily project<sup>17</sup>.

---

<sup>16</sup> [<http://ancientgraffiti.wlu.edu/>] (Benefiel, Sypniewski, & Sprenkle, 2017).

<sup>17</sup> [<http://sicily.classics.ox.ac.uk/>] (Prag, 2017). See Chapter 19 in this volume.

Adnuntiatio		la
Translated term	munera	la
Definition	Bekanntmachung; Ankündigung (z. B. von munera) Nicht Rechtliche Verfügungen (s.Rechtliche Verfügung, öffentlich / privat)	de
Examples	<a href="#">HD053407</a>	de
Created	2013-08-01 12:27:53	
Modified	2013-08-15 16:56:04	
agonistic / ludic		en
Translated term	agonistic / ludic	en
Translated term	agonistic / ludic	en
Created	2014-03-13 11:28:02	
Akklamation		de
Translated term	Acclamation	fr
Translated term	Aclamación	es

**Figure 17.1:** Main terms in different languages in the Type of Inscription EAGLE Vocabulary

The EpiDoc guidelines are, to date, a *de facto* standard for any new epigraphic project, digital or not, and have gone well beyond the limits of classical epigraphy, being used for a range of projects outside of the Mediterranean milieu (Elliott et al., 2007–2013).<sup>18</sup> The most interesting and rewarding feature of the encoding schema suggested is the encoding of the text, and this is the area where more effort has been focused. We had texts similar to the following (EDR000085) that needed to be converted to the EpiDoc XML with an automated process and then displayed on the portal with a unified edition style:<sup>19</sup>

-----  
[---? deposit]us in paç[e ---?],  
[cons(ulatu) Nicomaci Fl]abiani v̄(iri) [c(larissimi)].

<sup>18</sup> An idea of the number of projects is given by this category page on the Digital Classicist wiki [<http://wiki.digitalclassicist.org/Category:EpiDoc>]. See also Chapters 1 and 6 in this volume.

<sup>19</sup> [[http://www.edr-edr.it/edr\\_programmi/res\\_complex\\_comune.php?id\\_nr=EDR000085](http://www.edr-edr.it/edr_programmi/res_complex_comune.php?id_nr=EDR000085)].



The rudimental up-conversion<sup>20</sup> stylesheets used returned this result:

```
<ab>
  <lb n="0"/><gap reason="lost" extent="unknown" unit="line"/>
  <lb n="1"/><gap reason="lost" extent="unknown" unit="character">
    <certainty locus="name" match=".." cert="low"/></gap>
    <supplied reason="lost"> deposit</supplied>us in pa<unclear>c</unclear><supplied reason="lost">e
  </supplied><gap reason="lost" extent="unknown" unit="character"><certainty locus="name" match=".."
    cert="low"/></gap>,
  <lb n="2"/><supplied reason="lost"><expan><abbr>cons</abbr><ex>ulatu</ex></expan> Nicomaci
  Fl</supplied><choice><sic><unclear>a</unclear>biani</sic><corr>Flaviani</corr></choice> v̄(iri)
  <supplied reason="lost"><expan><abbr>c</abbr><ex>larissimi</ex></expan> </supplied>.
</ab>
```

This is then visible on the portal as follows:

[---] / [---? deposit]us in paç[e ---?], / [cons(ulatu) Nicomaci  
Fl] /Flaviani<sup>1</sup> v̄(iri) [c(larissimi) ].

You will immediately notice where the problems are:

1. the abbreviation v̄(iri) is not matched,
2. the corrected version of the word *flaviani* instead of *flabiani*, as on the stone, is wrongly encoded; the <supplied> element should have been split to have *Fl* inside the element <sic> of the <choice> and then unified in the visualization with the previous <supplied> element,
3. an unwanted space appears in the last portion of text supplied by the editor after the expansion of the abbreviation for *clarissimi*.

The user would have noticed almost nothing on the portal if we had decided to show the content of the element <sic> instead of <corr>. He/she still does not see the error for the first problem, v̄(iri), which remains untouched, as in the source, by the XSLT rendering the text.

This is clarified in the portal to guide the users, but it remains a problem to be resolved by improving the algorithms for the up-conversion, or fixing by hand where needed.

However, too many hands would be needed for more than 500,000 inscriptions. Therefore, the first solution and especially the second need to be implemented collaboratively (Feraudi-Gruénais & Grieshaber, 2016).

It was not only data from various types of databases that had to be exported and mapped. EpiDoc data needed to be converted to the EAGLE EpiDoc. Needless to say, it took orders of magnitude less time and effort to do this, and in these cases no text up-conversion was needed and the correctness of the mark-up was guaranteed by the content provider.

<sup>20</sup> The proceeding of the latest Balisage Markup Conference are very instructive on this topic (*Proceedings of Balisage: The Markup Conference 2017*, 2017).

Still, there are other problems related to the time available for the transformation, which lead to inconsistency in the data display. One could test this, which is fortunately “only” a visualization problem for correct underlying data, searching for one of the Roman Inscriptions of Britain, which often have three parallel editions in the EAGLE data.

Let us look for example at TM 154498, which is present in RIB, EDCS and EDH.

The RIB 5<sup>21</sup> XML for the text looks like this:

```
<div type="edition" xml:lang="la" xml:space="preserve">
  <ab type="original">Num(ini) C[aes(aris) Aug(usti)] | prov[incia] | Brita[nnia]</ab>
  <ab type="markup">
    <lb n="1"/><persName key="db04709" type="divine"><w lemma="numen"><expan>Num<ex>ini</ex>
  </expan></w> <name nymRef="#Caesar">C<supplied reason="lost"><expan>aes<ex>aris</ex></expan></supplied></name>
  <name nymRef="#Augustus"><supplied reason="lost"><expan>Aug<ex>usti</ex></expan></supplied></name>
  </persName>
    <lb n="2"/><w lemma="provincia">prov<supplied reason="lost">incia</supplied></w>

  <lb n="3"/><region key="db04957" type="province"><name nymRef="#Britannia">Brita<supplied reason="lost">nnia</supplied></name></region>
  </ab>
</div>
<div type="translation" xml:lang="en" xml:space="preserve">
  <ab>To the <persName key="db04709" type="divine">Divinity of the Emperor</persName> the province of <region key="db04957" type="province">Britain</region> (set this up).</ab>
</div>
```

EDCS-07800230<sup>22</sup> after transformation and up-conversion has the following XML, which is equal to that independently produced exporting HD069342<sup>23</sup>:

```
<div type="edition" xml:lang="la">
<head>Text</head>
<ab>
  <lb n="1"/>
  <expan><abbr>Num</abbr><ex>ini</ex></expan> C<supplied reason="lost"><expan><abbr>aes</abbr><ex>aris</ex>
  </expan> <expan><abbr>Aug</abbr><ex>usti</ex></expan></supplied>
  <lb n="2"/>prov<supplied reason="lost">incia</supplied> <lb n="3"/>Brita<supplied reason="lost">nnia</supplied>
</ab>
</div>
```

The visualization of RIB on the portal to date apparently takes both the `<ab>` inside `<div type="edition">` as well as the translation.

Given the volume of data and issues that we encountered during the project, several minor issues still remain to be resolved. This will happen in the very near future, once a particular member of the association, with knowledge and access to the data, is available to solve them. If the application and the data were collectively maintained, this would not need to wait so long.

Specificities in encoding are fine, and EpiDoc does a great job of allowing enough rigour and enough flexibility, thus serving perfectly its aim and its diverse users. Still

21 [https://romaninscriptionsofbritain.org/inscriptions/5].

22 [http://db.edcs.eu/epigr/edcs\_id.php?s\_sprache=en&p\_edcs\_id=EDCS-07800230].

23 [http://edh-www.adw.uni-heidelberg.de/edh/inschrift/HD069342].

we must be aware it is not enough alone and does not do magic just because is EpiDoc. We need to do more and better EpiDoc, and to keep training people and making it a research quality point for new students and scholars.

## 17.4 Methodological Issues Faced After EAGLE

We tried never to say that the mappings and conversions were perfect.<sup>24</sup> They are not, and still they serve a great function and lead the way for much more. Let us list some methodologically critical points in the harmonization process once more:

- some partners did not provide EpiDoc at all and preferred to deliver the data with other formats (their data has never been up-converted),
- some partners sent their data too late and the up-conversion could not be made precise enough,
- some partners do not strictly follow data entry guidelines and the up-conversion process fails more often than it should, relying on consistency,<sup>25</sup>
- the export and transformation workflow will always need checking and updating, thus it is not sustainable as a workflow,<sup>26</sup>
- some datasets have TM IDs and some not,
- some datasets add to some elements references to Trismegistos GEO IDs, some others do not,
- the editing of the vocabularies follows a GitHub based workflow which is efficient but not particularly user friendly.<sup>27</sup>

Let us take as an example a task that sounds easy.<sup>28</sup> Let us try to extract EAGLE data about the provinces relevant for a specific project like LatinNow.<sup>29</sup> Besides temporary issues of the portal, where sometimes the actual XML cannot be downloaded and the function to export results only saves the one in the current view, the results obtained with any general search would have been partial. This is the case with any database, as the user is constrained by the provided functionality. But in this case I could easily,

---

**24** On data flow quality (Mannocci, 2017).

**25** A test analysis of the cleanliness of data was done in 2014 for EDH, with astonishingly good results and a very high level of cleanness in the data entered.

**26** The CNR-ISTI built a very useful Content Checker, unfortunately very little used by the partners, which should be valorised in the future.

**27** An accessible editor integrated in the workflow is missing although many exist and one was developed within EAGLE as well by the CNR-ISTI team.

**28** All data analysis has been carried out using XQuery in a local version of exist-db 3.5 [<http://exist-db.org>].

**29** ERC project LatinNow (Latinization of the north-western Roman provinces: sociolinguistics, epigraphy and archaeology, grant number 715626) [<https://latinnow.eu/>], PI Alex Mullen.

thanks to the IDEA association, access all the EAGLE data to provide a better answer and deliver the required data.<sup>30</sup>

The data from the content providers uses different definitions of provinces, depending, for example, on the time scope of the original database or on internal definitions.

We must first isolate the data belonging to the selected provinces. In the EAGLE EpiDoc model this is information stored in a TEI element `<placeName type="provinceItalicRegion">` which, in the expectations of users, should be indexed by the aggregator to provide a filter “by province” and provide the functionality to search with this criterion, even if different denominations have been used, i.e. avoiding the bare string matching. The assumption is that the field is aligned to a Trismegistos GEO ID and that this is used as a key to group different denominations (Evangelisti, Liuzzo, & Verreth, 2014; Verreth, 2017). For example, by the documentation I would expect an inscription from Lusitania to have the following tag:

```

<placeName
  ref="http://www.trismegistos.org/place/5531"
  type="provinceItalicRegion"
  Lusitania
</placeName>

```

However, working directly with the current raw data in EAGLE, the values of this element are quite different. Out of 412,757 document entities in the dataset with this `<placeName>` tag (i.e. almost 100k do not have any), only 77,303 have a `@ref` pointing to the Trismegistos GEO ID.<sup>31</sup> There is also some expected ‘dirtiness’, like some data with `<placeName type="provincItalicRegion">` instead of the correct value of the attribute `@type`. For this reason, no filter by province is offered. The results would be more imprecise than searching for Lusitania, in the place of provenance. Actually, querying the data directly, there are 277 different values for this element and for example, “Narbonensis” appears in Gallia Narbonensis, Narbonensis, Narbonensis?, Narbonensis II, Narbonensis I. This is an acceptable workaround for the website, as it is intuitive without forcing high expectations. The user knows that the portal is an aggregator of heterogeneous data and will most often use this parameter, together with others, to run not one search but several. Since the volume of aligned entities

---

<sup>30</sup> The observations made here are based on data downloaded 2017/10/31. Many thanks go especially to Claudio Atzori, Andrea Mannocci and Franco Zoppi at CNR-ISTI in Pisa who have answered my requests faster than one could ever expect.

<sup>31</sup> Many more have this for the precise find-spot instead, which allows us to offer in the website the ancient find-spot filter with a decent degree of reliability. Of these 77,303 almost all are records from the Epigraphic Database Heidelberg.

is not sensible compared to the corpus, one must take into account the diversity of values and group-values that probably belong to the same province of interest by hand.

Once we have all document entities referring to one of the values for the province (thus reasonably all the entities referring to the desired province), these need to be grouped by TM ID to have unique texts and their multiple editions. This is possible only to the extent to which there is such information in the data, and that it is updated and correct; this is not easy. In the dataset used for this paper, 391,227 documental entities of a total of 502,961 have at least one. The EAGLE aggregator can attribute many more on the basis of the updated Trismegistos IDs, even without injecting this information in the source data. Some content providers actually could not, during the life of the project, enter these IDs that became available later.

Trismegistos has accomplished the incredible task of disambiguating all existing digitized texts during the lifetime of the EAGLE project. However, the process of updating this information in the databases had to follow a procedure where the valuable correspondence tables were sent over from one partner to another. Within the scope of its action, IDEA has developed a small tool, based at the University of Hamburg, which serves this data via a data API. This tool can respond dynamically to the request for parallel texts connected to a Trismegistos ID, either starting the query from a local id, or from a Trismegistos ID and returning several common formats for developers to easily reuse the information in their applications.<sup>32</sup>

## 17.5 General Issues in Digital Epigraphy

There are currently several general issues in the field of epigraphic databases. I will list some and omit more general issues, e.g. the use of closed or private and inaccessible databases to provide results in publications and presentations, thus cutting out the verifiability of the results presented. This is a poor practice that we have observed in plenary presentations at international conferences. The following are just five selected points:

1. researchers who are not IT specialists, such as historians and philologists, are forced to traipse across an assortment of databases when seeking information about inscriptions, EAGLE being one of them in some cases, especially thinking of the lack of Greek texts;
2. the wealth of information inscribed within texts, the connections between text, support and context have been discussed extensively but are still largely

---

<sup>32</sup> Since this article has been submitted a major improvement has taken place, as this service in a much better and richer way is provided directly by Trismegistos texrelations API at [<https://www.trismegistos.org/dataservices/texrelations/documentation/>].

underexploited in print, where little can be done about it, but also in digital resources where these connections when explicit could be easily and fruitfully used;<sup>33</sup>

3. resources in attested languages of which the researcher is not aware become blind spots, which is in contradiction with the multilingual nature of societies of the past;
4. crucially, these resources fail to be adequately referenced and used across publications in the epigraphic realm;
5. only authors and editorial teams can directly contribute, all others have to take more or less complicated workarounds.

Some other, more specific issues could be listed for projects like EAGLE, where the aim of aggregating data for Europeana has forced some definitions at different levels (Liuzzo, 2015). However issues are not what should stop us, but rather should help us to progress. The interaction of different resources, a virtue of any discipline that no project should propose to obliterate, is a huge challenge, and problems in the processes such as those encountered are not surprising.

The first issue could be easily overcome through an aggregator such as EAGLE, if aggregation did not imply regular updates. These are not always possible, especially if a contributing project is discontinued or does not have the human resources to implement it.

The second issue has been only partially faced by EAGLE. The EAGLE vocabularies and the partial EpiDoc encoding of the text go in this direction, allowing the visualization of related results based on one of the aligned features, but the automated mark-up needs to be edited and can only be considered a facilitator for the beginning of a real digital edition, rather than the final product of a process. Existing projects requesting EAGLE data for other purposes are the ideal user of this data and have been numerous. The EPNet project<sup>34</sup> is using the very promising federated databases approach (Calvanese et al., 2016) and the CRMtex group is also making a very positive effort for the creation of a CIDOC-CRM model for epigraphy (Felicetti & Murano, 2016; Ruiz, Vassallo, & Liuzzo, 2014).

The third issue is more interesting because it is an issue of the discipline, not only of digital resources, which can be really supported by new digital resources thus serving not just the immediate needs of current research, but opening up an entirely new set of questions and possibilities for it. Beside the examples of the CIIP (Cotton et al., 2010) and I.Sicily a comprehensive Epigraphy, without further labels,

---

**33** Only EAGLE has, to my knowledge a rudimentary attempt to show significantly related resources. The IDEs project [<https://blogs.library.duke.edu/dctthree/projects/>] is instead doing this in a much cleverer way for Greek inscriptions.

**34** [<http://www.roman-ep.net/wb/2016/12/22/ceipac-database-updated/>].

has never existed, not to speak of any comparative effort, and could in fact not exist until this days when it can be leveraged by proper tools.<sup>35</sup> Digital tools based on properly curated and linked data can help the researcher on these points.

The fourth problem is again one outside the strict realm of digital epigraphy, but affects all digital resources. Why should a contribution to an online resource which everyone uses and reads not be evaluated and accounted for in the evaluation of the scientific activity of a researcher as a paper, when this is also properly peer reviewed? There is here a hole in the more general system, but also digital resources have not done their part to make it possible and easy, although it could have been easier than thought. Now there is really no more excuse for researchers not to properly cite digital resources, as there is no excuse for digital resources not being easily and readily citable. Nevertheless, it is very rare to find the precise citation of digital resources in papers, as it is difficult to find the proper method to cite a digital resource. To make the citation of epigraphic digital resources possible, will be one of the central scopes of the already mentioned editor & navigator Epigraphy.info (Feraudi-Gruénais & Grieshaber, 2016).<sup>36</sup> The point of the evaluation of such research products needs to be discussed in the proper venues and certainly requires far-sighted advocates.

The last issue highlighted here, is dependent on the previous one and requires the biggest leap of faith: opening one's own editorial work to the contributions of others. Assuming we start thinking of digital editions as critical editions, we edit them as such and we offer them to the public as such, then we need a further step to make them editable by others.

The work carried out to facilitate digital publication has also received, in the last year, a major input with the release of EFES (EpiDoc Front-End Services),<sup>37</sup> EVT 2 beta 1 (Edition Visualization Technology)<sup>38</sup> and TEI-Publisher for exist-db.<sup>39</sup> This last, which I have personally tested, allows direct publication of TEI files in a way that has never been so easy (Turska, Cummings, & Rahtz, 2016; Wicentowski & Meier, 2015). It is usable out-of-the-box for TEI Simple, but it is also very easy to use with the EpiDoc ODD.

---

<sup>35</sup> Only Trismegistos to my knowledge does host and integrate data in all different languages.

<sup>36</sup> [<http://epigraphy.info/>].

<sup>37</sup> [<https://github.com/EpiDoc/EFES>].

<sup>38</sup> [<http://evt.labcd.unipi.it/>].

<sup>39</sup> [<http://teipublisher.com/index.html>].

## 17.6 Conclusions

Whilst it has never been possible to directly enrich a specific dataset with data from other datasets, no comparative approach has ever been served by a digital resource for inscriptions either. These would greatly enhance the range of possible research questions that could be addressed. Research has always remained within linguistic, chronological and spatial boundaries that EAGLE, for the first time, attempted to overcome, hosting inscriptions in all languages. In addition, it is to be noted that epigraphic research lacks entirely, not just digitally, a viable way to view the current status of digitization. Instead, some online resources are happy with giving the false impression that “everything” is already there, thus building a chain of misunderstandings, leading to the misuse of online resources. Few disciplines can be as proud of having so many texts online as classical epigraphy. For even fewer, it would make more sense to have a common overview of who is doing what and where and to ensure that the increasingly limited resources are not wasted in the repetition of tasks, whilst other research areas remain forever untouched.

## Bibliography

- Amato, G., Bollettieri, P., Gennaro, C., Manghi, P., Mannocci, A., & Zoppi F. (2013). *AIM Infrastructure Specification*. [Report of the project EAGLE]. Retrieved from [[https://www.eagle-network.eu/wp-content/uploads/2013/06/EAGLE\\_D4.1\\_AIM\\_Infrastructure\\_Specification\\_update.pdf](https://www.eagle-network.eu/wp-content/uploads/2013/06/EAGLE_D4.1_AIM_Infrastructure_Specification_update.pdf)], 2017/10/31.
- Benefiel, R., Sypniewski, H., & Sprenkle, S. (2017). Working with Text and Images: The Graffiti of Herculaneum. In S. Orlandi, R. Santucci, F. Mambrini, & P.M. Liuzzo, (Eds.), *Digital and Traditional Epigraphy in Context. Proceedings of the EAGLE 2016 International Conference* (pp. 145–159). Roma: Sapienza Università Editrice. doi: 10.13133/978-88-9377-021-7
- Bigi, F. (2014). Towards an EAGLE Standard in Translating Inscriptions. In S. Orlandi, R. Santucci, V. Casarosa, & P.M. Liuzzo (Eds.), *Information Technologies for Epigraphy and Cultural Heritage: Proceedings of the First EAGLE International Conference* (Serie antichistica. Collana Convegni 26) (pp. 167–178). Roma: Sapienza Università Editrice. Retrieved from [<https://www.eagle-network.eu/wp-content/uploads/2015/01/Paris-Conference-Proceedings.pdf>], 2017/11/30.
- Calvanese, D., Liuzzo, P., Mosca, A., Remesal, J., Rezk, M., & Rull, G. (2016). Ontology-based data integration in EPNet: Production and distribution of food during the Roman Empire. *Engineering Applications of Artificial Intelligence*, 51, 212–229. doi: 10.1016/j.engappai.2016.01.005
- Cotton, H.M., Segni, L.D., Eck, W., Isaac, B., Price, J., Kushnir-Stein, A., Misgav, H., Price, J., Roll, I., & Yardeni, A. (2010). *Corpus Inscriptionum Iudaeae / Palaestinae: A multi-lingual corpus of the inscriptions from Alexander to Muhammad* (1<sup>st</sup> ed., Vol. 1.1). Berlin/New York: De Gruyter.
- Elliott, T., Bodard, G., Milonas, E., Stoyanova, S., Tupman, C., & Vanderbilt, S. (2007–2013). *EpiDoc Guidelines: Ancient documents in TEI XML*. Retrieved from [<http://www.stoa.org/epidoc/gl/latest/>], 2017/12/09.
- Evangelisti, S., Liuzzo, P.M., & Verreth, H. (2014). *Content Harmonisation guidelines, including GIS and terminologies*. [Deliverable of the project EAGLE, 1st release]. Retrieved from



- [[https://www.eagle-network.eu/wp-content/uploads/2013/06/EAGLE\\_D2.2.1\\_Content-harmonisation-guidelines-including-GIS-and-terminologies.pdf](https://www.eagle-network.eu/wp-content/uploads/2013/06/EAGLE_D2.2.1_Content-harmonisation-guidelines-including-GIS-and-terminologies.pdf)], 2017/12/09.
- Felicetti, A. & Murano, F. (2016). Scripta manent: a CIDOC CRM semiotic reading of ancient texts. *International Journal on Digital Libraries*, 18(4), 263–270. doi: 10.1007/s00799-016-0189-z
- Feraudi-Gruénais, F. & Grieshaber, F. (2016). Digital Epigraphy am Scheideweg? / Digital Epigraphy at a crossroads? [Presented at the Nachnutzung und Nutzbarkeit der Forschung im Akademienprogramm Workshop der Nordrhein-Westfälischen Akademie der Wissenschaften und der Künste und der Union der deutschen Akademien der Wissenschaften AG „eHumanities“, Düsseldorf]. doi: 10.11588/heidok.00022141
- Liuzzo, P.M. (2015). EAGLE and EUROPEANA. Architecture Problems for Aggregation and Harmonization. In *Proceedings of the Symposium on Cultural Heritage Markup* (Balisage Series on Markup Technologies 16). Retrieved from [<http://www.balisage.net/Proceedings/vol16/html/Liuzzo01/BalisageVol16-Liuzzo01.html>], 2017/12/09.
- Liuzzo, P.M. (2017). Mapping Databases to EpiDoc. In S. Orlandi, R. Santucci, F. Mambrini, & P.M. Liuzzo, (Eds.), *Digital and Traditional Epigraphy in Context. Proceedings of the EAGLE 2016 International Conference* (pp. 187–200). Roma: Sapienza Università Editrice. doi: 10.13133/978-88-9377-021-7
- Liuzzo, P.M., Fasolini, D., & Rocco, A. (2014). *Content Harmonisation guidelines, including GIS and terminologies*. [Deliverable of the project EAGLE, 2nd release]. Retrieved from [[https://www.eagle-network.eu/wp-content/uploads/2013/06/EAGLE\\_D2.2.2\\_Content-harmonisation-guidelines-including-GIS-and-terminologies-Second-Release.pdf](https://www.eagle-network.eu/wp-content/uploads/2013/06/EAGLE_D2.2.2_Content-harmonisation-guidelines-including-GIS-and-terminologies-Second-Release.pdf)], 2017/12/09.
- Liuzzo, P.M., Mambrini, F., & Franck, P. (2017). Storytelling and Digital Epigraphy-Based Narratives in Linked Open Data. In M. Ioannides, N. Magnenat-Thalmann, & G. Papagiannakis (Eds.), *Mixed Reality and Gamification for Cultural Heritage* (pp. 507–523). Springer: Cham. doi: 10.1007/978-3-319-49607-8\_20
- Manghi, P., Mannocci, A., Sicilia, M.A., Gomez Pantoja, J., Rubiro Fuentes, J., Rivero Ruiz, E., & Zoppi, F. (2015). *EAGLE metadata model specification*. [Deliverable of the project EAGLE, 2nd release]. Retrieved from [[https://www.eagle-network.eu/wp-content/uploads/2013/06/EAGLE\\_D3.1\\_EAGLE-metadata-model-specification\\_v1.1.pdf](https://www.eagle-network.eu/wp-content/uploads/2013/06/EAGLE_D3.1_EAGLE-metadata-model-specification_v1.1.pdf)], 2017/12/09.
- Mannocci, A. (2017, January 12). *Data Flow Quality Monitoring in Data Infrastructures* (PhD thesis). Pisa: Università di Pisa. Retrieved from [[https://etd.adm.unipi.it/theses/available/etd-12232016-151401/unrestricted/PhD\\_Thesis.pdf](https://etd.adm.unipi.it/theses/available/etd-12232016-151401/unrestricted/PhD_Thesis.pdf)], 2017/12/09.
- Orlandi, S., Santucci, R., Casarosa, V., & Liuzzo, P.M. (Eds.). (2014). *Information Technologies for Epigraphy and Cultural Heritage: Proceedings of the First EAGLE International Conference* (Serie antichistica. Collana Convegna 26). Roma: Sapienza Università Editrice. Retrieved from [<http://archiv.ub.uni-heidelberg.de/propylaeumdok/2337/>], 2017/12/09.
- Orlandi, S., Santucci, R., Mambrini, F., & Liuzzo, P.M. (Eds.). (2017). *Digital and Traditional Epigraphy in Context. Proceedings of the EAGLE 2016 International Conference*. Roma: Sapienza Università Editrice. doi: 10.13133/978-88-9377-021-7
- Prag, J. (2017). I.Sicily: an epidoc corpus for ancient Sicily. In S. Orlandi, R. Santucci, F. Mambrini, & P.M. Liuzzo, (Eds.), *Digital and Traditional Epigraphy in Context. Proceedings of the EAGLE 2016 International Conference* (pp. 83–96). Roma: Sapienza Università Editrice. doi: 10.13133/978-88-9377-021-7
- Proceedings of Balisage: The Markup Conference 2017*. (2017). (Balisage Series on Markup Technologies 19). Mulberry Technologies, Inc. Retrieved from [<https://www.balisage.net/Proceedings/vol19/masthead.html>], 2017/12/09.
- Ruiz, E.R., Vassallo, V., & Liuzzo, P.M. (2014). Networking EAGLE with CIDOC and TEI. [Presented at Conference: CIDOC 2014: Access and Understanding – Networking in the Digital Era]. Retrieved from [[http://www.cidoc2014.de/images/sampleddata/cidoc/papers/1-2\\_Vassallo\\_Ruiz\\_Liuzzo\\_paper.pdf](http://www.cidoc2014.de/images/sampleddata/cidoc/papers/1-2_Vassallo_Ruiz_Liuzzo_paper.pdf)], 2017/12/09.

- Turska, M., Cummings, J., & Rahtz, S. (2016). Challenging the Myth of Presentation in Digital Editions. *Journal of the Text Encoding Initiative*, 9. doi: 10.4000/jtei.1453
- Verreth, H. (2017). Trismegistos Places, a geographical index for all Latin inscriptions. In S. Orlandi, R. Santucci, F. Mambrini, & P.M. Liuzzo, (Eds.), *Digital and Traditional Epigraphy in Context. Proceedings of the EAGLE 2016 International Conference* (pp. 201–208). Roma: Sapienza Università Editrice. doi: 10.13133/978-88-9377-021-7
- Wicentowski, J.C. & Meier, W. (2015). Publishing TEI documents with TEI Simple. In *Proceedings of Balisage: The Markup Conference 2015* (Balisage Series on Markup Technologies 15). doi: 10.4242/BalisageVol15.Wicentowski01

Thomas Kollatz

## 18 EPIDAT – Research Platform for Jewish Epigraphy

**Abstract:** EPIDAT, the research platform for Jewish epigraphy, deals with Jewish epigraphy in all its aspects. This article describes the on-going project and data-driven development, since the year 2002, which resulted in a wide range of access options to the epigraphic records. Later on, the solid data basis hosted by EPIDAT enabled cooperation across disciplines (linguistics, art history, monument science, cultural heritage agencies) and other epigraphic projects. Interoperability is essential for epigraphy, but needs reliable ontologies and cooperation over several projects and beyond disciplines.

**Keywords:** data analysis, interoperability, Jewish studies, semantic web, visualisation

### 18.1 Introduction

“If there were a database containing as many inscriptions as possible from a huge number of cemeteries in Germany as well as Central and East Europe, we would then have a corpus of source material appropriate for many research questions; both known research questions as well as new ones – inter alia the study of the differences between eulogies for women and men” (Brocke & Mirbach, 1988).

It would take almost 20 years before a database for Jewish epigraphy was established. In 2002 the Salomon Ludwig Steinheim-Institute for German-Jewish History (then located in Duisburg, now Essen) was commissioned to carry out the photographic and scientific documentation of the Ashkenazic Cemetery of the Jewish Communities of Hamburg and Altona. It was clear from the beginning that this task could only be effectively and adequately handled with the aid of computers. In order to present them in their entirety, more than 6,000 remaining objects required a digital edition. “It was [...] the sheer question of volume which led [...] to start considering electronic publication” (Roueché, 2010). A team of Hebrew and Jewish Studies specialists, together with art-historians and monument conservators defined the base structure, the principal set-up and the main categories needed in order to enable adequate research on the headstones preserved on this extraordinary cemetery. Even though the initial aim was to provide a straightforward method for accessing the rich

---

**Thomas Kollatz**, Salomon Ludwig Steinheim-Institute for German-Jewish History, Essen; Academy of Science and Literature, Mainz



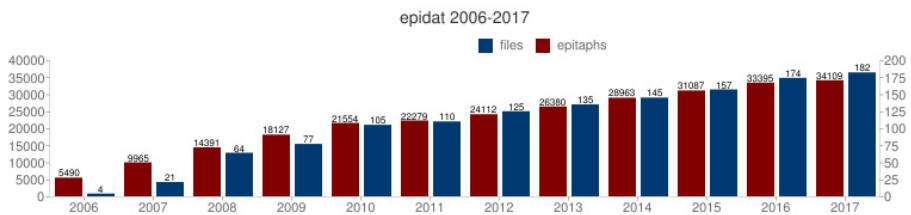
© 2018 Thomas Kollatz

This work is licensed under the Creative Commons Attribution 4.0 International license (CC BY 4.0)

epigraphical collection, it soon became clear that a structured database would have been also a superior tool for all kind of current and future research. In 2006 EPIDAT was officially launched on the Internet.<sup>1</sup>

## 18.2 EPIDAT Metadata Collections

Based on several epigraphic projects in line with the Hamburg project, the database could constantly be developed and maintained (Kollatz, 2015).



**Figure 18.1:** Epitaphs and files (*per anno*)

Currently it contains the digital collections of about 180 historic Jewish cemeteries, with the edition of more than 33,000 epitaphs and 65,000 image files (Figure 18.1). We are not dealing with “big data”, but, in a sense, “long data” (Arbesman, 2013/01/29) or “small data” (Pollok, 2013/04/22) could be considered as a more proper term: the geographical focus is on Germany, but inscriptions from Jewish cemeteries in The Netherlands, and recently also from Lithuania and the Czech Republic are also collected. The time span ranges from the 11<sup>th</sup> to the 20<sup>th</sup> century, from the Medieval and Early Modern periods to the Modern Era.

The large spatial and chronological distribution of the resources requires a broad range of access points into the collections. Therefore, the epigraphical sources can be browsed through a number of filters that can be grouped as follows:

- Location-based filters: for users whose research interests focus on the epigraphical tradition of a single community or a certain region.
- Time-based filters for research into textual and visual features changing over time.

Epigraphs, both in their physical and in the textual aspects, are indexed. Indexes concern: symbols depicted on the headstones, word forms used in the inscriptions as well as quotations from the Hebrew Bible, references to rabbinic literature and

<sup>1</sup> [<http://www.steinheim-institut.de/cgi-bin/epidat>].

liturgical books, persons mentioned on the inscriptions, and stonemasons involved in the craft, etc. Moreover, a full-text search helps finding keywords and idiomatic expressions.

A huge number of digital images can be browsed independently for chronology and by provenance.

Different kinds of maps show where any single cemetery is located as well as regions wherein a large number of cemeteries could be documented.

### 18.3 Text Encoding

In order to promote and to encourage reuse of EPIDAT-records, a machine-readable open interface has been made available. This web interface ensures that the epigraphical datasets are harvested and downloaded by third parties. Since 2008 EPIDAT records are provided in EpiDoc.<sup>2</sup>

A special opportunity was the cooperation between the EPIDAT team and building researchers, made possible by funding from the Federal Ministry of Education and Research (BMBF). The cooperation between “text-minded” and “object-minded” researchers proved to be useful and broadening. It appeared to be instructive to pay attention to the text-bearing objects themselves. It is not only the text, where “it” happens: cultural change is not only expressed by textual means, especially with respect to epitaphs, which by nature are a conservative and traditional medium. When text is fixed and subject to conventions, then the form of the objects could be the vehicle for change in religion, culture and society (Figure 18.2).

In the scope of the cooperative project, a preliminary object mark-up schema was developed that enabled us to merge textual and object-related data (Arera-Rütenik & Kollatz, 2016). For this kind of transdisciplinary research perhaps, in the future, an “Object Encoding Initiative” could meet the requirements of less text-orientated disciplines such as art history, history of architecture, iconology and visual analytics. Digital epigraphy should take into account the methodological requirements of all involved disciplines. As far as the encoding of text is concerned, EpiDoc is a stable basis for all kinds of text-orientated approaches in philology, religious and cultural studies. The picture is different when it comes to encoding objects, which are more than just text-bearing objects:

“What characterises this class of objects is that they form a whole with their physical support. Indeed, the meaning of an epigraph cannot be fully understood without the analysis of the object or monument or other archaeological object on which it appears, just as one cannot fully understand the nature of that particular archaeological object without thoroughly investigating the sense of the inscription or iconographic representation it hosts” (Felicetti et al., 2016).

---

<sup>2</sup> [<http://www.stoa.org/epidoc/gl/latest/toc-it.html>].

**ALTER JÜDISCHER FRIEDHOF IN BONN-SCHWARZREINDORF**

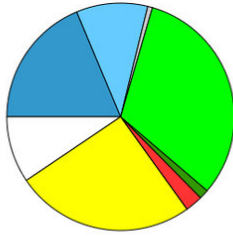
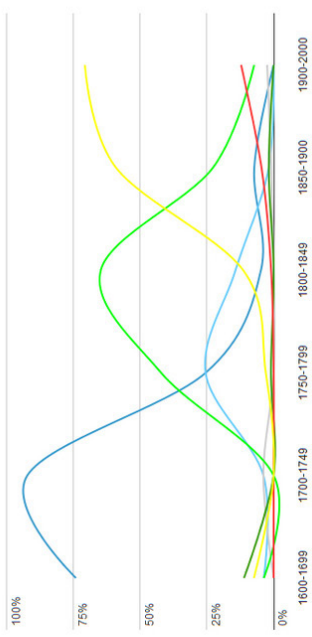
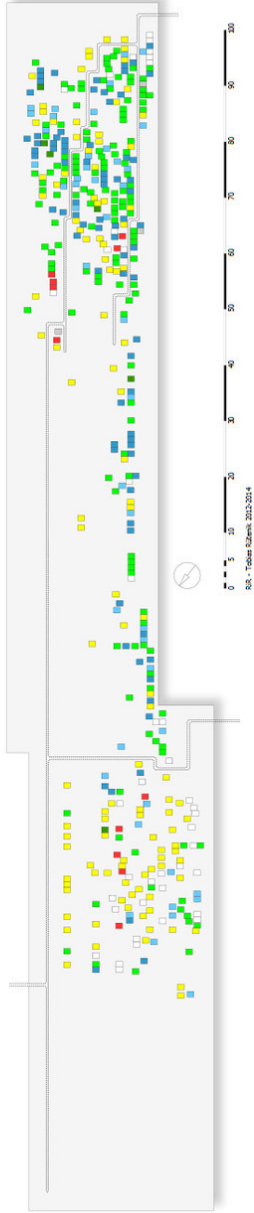
**obere Abschlüsse in Kubatur:**

- Rundbogen - eingezogen
- Segment-/Korbbogen - eingezogen

- andere - eingezogen
- Karmelbogen

- gegeneinander gest. Voluten
- Pyramide

- andere - nicht eingezogen
- nicht vorhanden/nicht sichtbar



**Figure 18.2:** Development of upper part forms of headstones

Epigraphy is an ancillary science and, indeed, it serves the vital purpose of the scientific community at large. Each and every object opened up and dealt with using epigraphic science is a unique source. Four specific features of the genre of the (Jewish) epitaph should not be underestimated:

- first, the fact that the majority of the objects can be precisely dated – by the date of death of the deceased person mentioned in the eulogy;
- secondly, the fact that the records can be located – by the place of burial;
- third, the fact that they could be clearly assigned to a gender – the headstone is erected to remember a certain man or woman, the eulogy reflects on her or his life, moreover it reflects common values a community of a certain period and a certain place have attributed to men and women;
- last, but not least, there is the sheer quantity of relevant data.

The stable quality of data and sufficient quantity of temporal, spatial and gender related information contributed by epigraphy, is suitable for a wide area of inter- and trans-disciplinary research questions. The names mentioned are a valuable source for genealogy and onomastic studies. The use of quotations from traditional literature in the eulogies testifies to the temporal preferences and spirit of a period. The same is true for the symbols shown on the headstones, and the materials and techniques used. The above-mentioned features can potentially contribute to both lexicography and linguistics.

Recently, EPIDAT provides in a beta-version the list of word forms sorted according to various criteria: words only occurring on inscriptions for women, occurrence of words related to time, and length, where longer words indicate Hebraised place and person names, etc. (Grüntgens & Kollatz, 2018). A problem, which is currently still unsolved, is that we are still missing part-of-speech taggers and lemmatizers for the Medieval and Early Modern periods. A lemmatized list would probably allow more refined conclusions on the development of the Hebrew language in the Diaspora and the impact of external non-Jewish culture.

## 18.4 Reuse of Data

Via the interface, EPIDAT records were actually more than once harvested and reused, e.g. in the scope of so-called hackathons. During the “Coding Da Vinci” event,<sup>3</sup> pushed forward by the Open Knowledge Foundation in order to make all kinds of cultural data available and known beyond the narrow framework of academic research, EPIDAT records were used by one of the project teams, and also during a pre-conference

---

<sup>3</sup> [<https://codingdavinci.de>].

workshop to the Digital Humanities Conference 2014 in Lausanne.<sup>4</sup> In addition, and perhaps just as important as these citizen science related activities, is the fact that the interface to EPIDAT allows and enables cooperation between disciplines.

EPIDAT metadata are also provided by a RSS-feed, not meant to describe every single headstone, but exclusively meant to provide information about the historical site and the history of the cemetery. These data are harvested on a regular basis by the mobile web application, “Places of Jewish History”, a web service developed by Harald Lordick, a researcher at the Steinheim-Institute. The mobile web application displays historical information on places near the user’s location, based on a wide range of relevant historical databases.

It is a remarkable fact that open and easy-to-use interfaces actually encourage all kinds of reuse. From a data curation perspective, it is equally important to enable traditional researchers to make good use of research data, as well as to ensure low-threshold access to the TEI-XML encoded research data. The former are learned readers, who could contribute to the quality and the stringency of the content by commenting upon, and discussing it. The latter are skilled users, who could assess the general rationale of the data structure by transferring and reusing it. Both the critical response on content as well as on the data model, its structure and form, is becoming increasingly important to the emerging digital humanities.

## 18.5 Interoperability

EPIDAT makes use of the infrastructure of the German digital library TextGrid<sup>5</sup> to enter and store data. TextGrid, in turn, adheres to the DARIAH network<sup>6</sup> and provides its collections to the aggregator. Both of them enhance the visibility of the EPIDAT collections. For instance, more than 20,000 dated inscriptions as well as about 3,000 dated headstones displaying symbols can be visualized in their mutual, spatio-temporal relations through the DARIAH Geo-Browser. This allows us to point out the geographical and chronological distribution of epigraphs recording particular names, or utilizing specific symbols.

Data output is available in different formats and schema: HTML5 (for the online digital edition), EDM (for the exposure to the Europeana harvester), KML (for spatio-temporal visualization), RTF and PDF (for printed publications), CSV for indexes, word lists, and recently in RDF to express formally the relations between persons named on headstones.

---

<sup>4</sup> [<https://dh2014.org>].

<sup>5</sup> [<https://textgridrep.org>].

<sup>6</sup> [<https://de.dariah.eu>].



All research data provided by EPIDAT are released online under an open Creative Commons license (CC-BY). The editorial principles of Open Data and Open Access are strictly observed.

Interoperability is of vital importance to draw attention to the potential impact Jewish epigraphy and its findings could have on Jewish studies, as well as on the humanities at large. In order to foster interoperability, EPIDAT records are provided and constantly enriched with metadata referring to controlled vocabularies, authority files, ontologies and thesauri. What we are still missing are robust ontologies that meet our disciplinary requirements. There are useful authority files for places and names;<sup>7</sup> however, the existing ontologies for art<sup>8</sup> and iconography<sup>9</sup> still lack appropriate categories for phenomena we come across in our specific domain.

We are aware that Jewish epigraphy is a marginal field of interest. A researcher interested in Hebrew poetry in the 17<sup>th</sup> century is usually unaware that hundreds of sophisticated poetic eulogies are preserved in contemporary Jewish cemeteries. The same holds for the symbols and ornaments, which are widely neglected by art and images sciences. Likewise, nobody would expect that headstones could contribute to the history of everyday things like “fish traps”, “ploughs”, or “shoes”. A particularity of medieval cities was that houses were referred to by symbols. A certain man lived in the house marked by the symbol “shoe”. From time to time this house name became a surname. It is remarkable that these house symbols are showing up on the headstones, also testifying to the development and change of everyday items over a long period (Figure 18.3).

The technical solution that could bring up all kinds of surprising findings of Jewish epigraphy to the surface is Semantic Web technologies. However, the way to the Linked Open Data cloud needs good preparation. Currently, we work together with neighbouring epigraphic projects, like Inscriptions of Israel and Palestine (IIP, Brown University)<sup>10</sup> and the recently begun Funerary Inscriptions of Jews from Italy (FIJI, Utrecht University) within Jewish Studies. However, we also work with the long-term project Deutsche Inschriften (German Inscriptions) (DI, Academies of Sciences in Germany and Austria) (Schrade, 2011). All projects mentioned are EpiDoc based. In a bilateral working group between EPIDAT and DIO (German Inscriptions Online, Digital Academy at Academy of Sciences and Literature Mainz)<sup>11</sup>, we have used the generic XTriples webservice, designed to extract semantic relations existing in XML resources (in our case: EpiDoc TEI XML). The service was developed in the

---

**7** For instance the CERL Thesaurus [[https://www.cerl.org/resources/cerl\\_thesaurus/main](https://www.cerl.org/resources/cerl_thesaurus/main)] or the Library of Congress authority files [<http://authorities.loc.gov>].

**8** Getty AAT [<http://www.getty.edu/research/tools/vocabularies/aat/>].

**9** Iconclass [<http://www.iconclass.nl/home>].

**10** [<http://cds.library.brown.edu/projects/Inscriptions/index.shtml>].

**11** [<http://www.inschriften.net>].

context of the long-term research project, *Deutsche Inschriften*, together with the project *Inscriptions in their Spatial Context* by Torsten Schrader, head of the Digital Humanities department of the Academy of Science and Literature Mainz (Grüntgens & Schrader, 2016). We succeeded in transforming TEI encoded family relations into RDF-statements as well mapping a complete corpus to CIDOC-CRM. The first results were promising and do lead us to rethink, evaluate and improve the shared data model, the TEI XML. Future plans are to provide structured data APIs, ideally a Sparql-Endpoint for both EPIDAT as well as German Inscriptions online.

In retrospect, in 2002 nobody expected EPIDAT to take such an evolution: what started as a practical workaround in order to handle one single historic cemetery, would develop into a research platform for Digital Jewish Epigraphy.



**Figure 18.3:** Buckled Shoe, Frankfurt 1795

## Bibliography

- Arbesman, S. (2013/01/29). Stop Hying Big Data and Start Paying Attention to ‘Long Data’, *Wired*. Retrieved from [<https://www.wired.com/2013/01/forget-big-data-think-long-data/>], 2017/11/30.
- Arera-Rütenik, T. & Kollatz, T. (2016). Interdisziplinäre Perspektiven auf Grabmale und Visualisierung räumlicher Strukturen. Ergebnisse eines Projektes zu historischen jüdischen Friedhöfen. In A. von Kienlin, K. Keßler, U. Knufinke, & S.M. Ross (Eds.), *Objekt und Schrift: Beiträge zur materiellen Kultur des Jüdischen* (Jüdisches Kulturerbe 1) (pp. 161–168). Braunschweig: TU Braunschweig.
- Brocke, M. & Mirbach, H. (1988). *Grenzsteine des Lebens*. Duisburg: Mercator.

- Felicetti, A., Murano, T., Ronzino, P., & Niccolucci, F. (2016). CIDOC CRM and epigraphy: A hermeneutic challenge. In P. Ronzino (Ed.), *Proceedings of the Workshop on Extending, Mapping and Focusing the CRM co-located with 19th International Conference on Theory and Practice of Digital Libraries (2015), Pozna, Poland, September 17, 2015*. Retrieved from [<http://ceur-ws.org/Vol-1656/paper5.pdf>], 2017/11/30.
- Grüntgens, M. & Kollatz, T. (2018). Korpusbasiertes Arbeiten und epigraphische Datenbanken: Möglichkeiten und Herausforderungen am Beispiel von Epidat und Dio. In J. Gessinger, A. Redder, & U. Schmitz (Eds.), *Korpuslinguistik (Osnabrücker Beiträge zur Sprachtheorie 92)* (pp. 157–174). Duisburg: Universitätsverlag Rhein-Ruhr.
- Grüntgens, M. & Schrade, T. (2016). Data repositories in the Humanities and the Semantic Web: Modelling, Linking, Visualising. In A. Adamou, E. Daga, & L. Isaksen (Eds.), *Proceedings of the 1st Workshop on Humanities in the Semantic Web co-located with 13th ESCW Conference 2016* (pp. 53–63). Anissaras. Retrieved from [<http://ceur-ws.org/Vol-1608/paper-07.pdf>], 2017/11/30.
- Kollatz, T. (2015). EPIDAT - Datenbank zur jüdischen Grabsteinepigraphik: Inventarisierung und Dokumentation historischer jüdischer Friedhöfe. In E. Bolenz, L. Franken, & D. Hänel (Eds.), *Wenn das Erbe in die Wolken kommt: Digitalisierung und kulturelles Erbe* (pp. 161–168). Essen: Klartext.
- Pollock, R. (2013/04/22). Forget Big Data, Small Data is the Real Revolution. *Open Knowledge International Blog*. Retrieved from [<https://blog.okfn.org/2013/04/22/forget-big-data-small-data-is-the-real-revolution/>], 2017/11/30.
- Roueché, Ch. (2010). Digitizing Inscribed Texts. In M. Deegan & K. Sutherland (Eds.), *Text Editing, Print and the Digital World* (pp. 159–168). Farnham: Ashgate.
- Schrade, T. (2011). Epigraphik im digitalen Umfeld, *Skriptum*, 1, 7–11. URN [urn:nbn:de:0289-2011051816]. Retrieved from [<http://www.skriptum-geschichte.de/2011/heft-1/epigraphik-im-digitalen-umfeld.html>], 2017/11/30.

Jonathan R.W. Prag and James Chartrand

## 19 I.Sicily: Building a Digital Corpus of the Inscriptions of Ancient Sicily

**Abstract:** This paper presents the I.Sicily project. We focus first upon its original rationale and construction, since this provides explanations for the particular choices and approaches adopted, before exploring some of the challenges faced, as well as current and future developments. We believe that I.Sicily offers an interesting case study of a deliberately open-ended, continuous work-in-progress corpus. The project is constructed on the assumption that collaboration is key to its success, and that collaboration will only increase. We examine the potential for the creation of Linked Open Data, which we consider essential to creating the primary point of reference for the study of Sicilian epigraphy, and to the creation of a resource to support and facilitate research while simultaneously enhancing and supporting the accessibility of Sicilian epigraphy. This last aim is served both directly through the project's web-interface, and indirectly by supporting and facilitating the work of the institutions which curate the majority of the material: we conclude with an illustration of a wide-ranging, museum-based, community collaboration.

**Keywords:** ancient Sicily, EpiDoc, museums, Linked Data, onomastics

### 19.1 Background

I.Sicily<sup>1</sup> is a corpus of the inscribed texts from ancient Sicily. This includes the very earliest written texts from the island (late seventh/early sixth century BCE), and extends to late Antiquity and the Byzantine period (seventh century CE). At present, for historical reasons and practical purposes, the primary coverage of the project is texts inscribed on stone (between 4,000 and 5,000 in total; currently 3,246 records). In due course, we will extend coverage to include other inscribed materials (especially metal and ceramic) and portable objects (*instrumentum domesticum*). A pilot project is under development to explore the creation of a sub-corpus of coin-legends in the same format. The epigraphic culture in ancient Sicily includes texts

---

1 I.Sicily [<http://sicily.classics.ox.ac.uk>].

---

**Jonathan R.W. Prag**, University of Oxford  
**James Chartrand**, Open Sky Solutions



written in Phoenician/Punic, Greek, Oscan, Latin, Hebrew, and two of the indigenous languages, Sikel and Elymian (for overviews of Sicilian epigraphy and linguistics, see the contributions in Gulletta, 1999; Tribulato, 2012a).

The original motivation for I.Sicily lies in traditional problems of publication and access. Sicily has a very long tradition of epigraphic study and corpus creation (De Vido, 1999): the first modern history of the island, which included epigraphic texts, is the *de rebus Siculis* of Tommaso Fazello (1558), and the first epigraphic corpus was published by Georg Gualtherus in 1624; Sicily was the subject of some of the earliest volumes of the monumental Berlin projects, *Corpus Inscriptionum Latinarum* (vol. X.2 = Mommsen, 1883) and *Inscriptiones Graecae* (vol. XIV = Kaibel, 1890). However, the rate of both discovery and publication increased rapidly from the late 1880s onwards, and the ability of both the primary publications (such as the gazettes, *Supplementum Epigraphicum Graecum* and *L'Année Épigraphique*) and scholars to keep pace with new material has been limited. The situation is compounded by the very uneven practices in the publication of archaeological excavation, and there is an unknown and not insignificant quantity of unpublished material (often highly fragmentary) languishing in stores across the island. Consequently, the discussion of Sicilian epigraphy has tended to be concentrated very narrowly in the hands of specialists, not simply for disciplinary reasons, but due to the difficulties of comprehensive knowledge (an emblematic example is Manganaro, 1988, an unparalleled discussion of the material of the Roman imperial period, referencing hundreds of texts, and alluding frequently to unpublished or obscure and unreferenced texts). These challenges have become even more visible in recent scholarship with the increased focus upon socio-linguistics, which depends upon the ability to engage with a comprehensive dataset. As Olga Tribulato recently noted, “Arguments [on the linguistic history of ancient Sicily], and the statistics on which they rely, are destined to remain little more than hypotheses, until a comprehensive list of all epigraphic texts from ancient Sicily is assembled” (Tribulato, 2012b, p. 324).

Against this backdrop, Jonathan Prag originally attempted to create just such a list of the lapidary inscriptions of Sicily. This was carried out within the framework of a PhD on Roman Sicily (London, 1999–2004), of which the initial results were published as a quantitative analysis (Prag, 2002), in order to assess epigraphic culture on the island. That project did not concentrate on the texts themselves, but on creating a reference list based upon bibliographic citations, together with a limited amount of metadata. The original list was created in a flat table in MS Access 97 (upgraded several times subsequently). This dataset was intermittently maintained and updated on a series of private computers over the following decade, during which time its value as a research tool became increasingly apparent.<sup>2</sup> The same period witnessed

---

<sup>2</sup> Facilitating e.g. Prag, 2003; 2007; 2008; 2010.

the development of the EpiDoc TEI-XML standard,<sup>3</sup> and in 2011 several bids were submitted to funding bodies to transform the existing dataset into an EpiDoc corpus. The primary funding for the creation of I.Sicily was provided by a grant of £80,000 from the John Fell Fund of the University of Oxford, which was used over the period 2013–2015.<sup>4</sup>

The principal development work undertaken over that period consisted of (a) the transformation of the legacy dataset from an Access table to a set of EpiDoc files; (b) the construction of the necessary back-end and front-end tools to make a usable corpus with a flexible web interface.<sup>5</sup> In its final form, the original table held data across some 40 different fields, for c. 3,200 records; 18 of these fields detailed publication history (corpora references and other bibliography); the other fields recorded information on the language, date, provenance, current location, epigraphic type, form and material of the inscriptions, together with a free-text field recording further information about the inscription and fields to record any autopsy undertaken. Almost all of this data was derived from existing publications. After extensive cleaning of the data, the conversion from the original MS Access dataset was developed through a pipeline of known conversions passing from MS Access to CSV to TEI P5 XML. The subsequent XSLT transformation of the table of data from TEI P5 XML to EpiDoc XML provided an ideal opportunity to enrich the existing dataset, both to normalise the data and to lay the foundations for Linked Open Data. This was done, principally, by the embedding of reference to multiple external authority lists (local correspondence lists were created in CSV files during the pre-conversion cleaning of the original data to facilitate this alignment). This enabled the incorporation of Pleiades and Geonames URIs on the “ref” attribute for `<placeName type="ancient">` and `<placeName type="modern">`, as well as the inclusion of representative decimal-degree location data in a `<geo>` element, to simplify local mapping. EAGLE vocabularies were incorporated for `@ref` on `<material>`, `<objectType>`, `<rs type="execution">` (in `<layout>`) and for epigraphic type on `<term>` within the `<textClass>` element.<sup>6</sup> Two new resources were created as part of the process of transforming the data: an open bibliography in Zotero<sup>7</sup> and a new museums database.<sup>8</sup> URIs are maintained for both sets of data (for bibliographic items

<sup>3</sup> [<https://sourceforge.net/p/epidoc/wiki/Home/>]; Bodard, 2010.

<sup>4</sup> Additional small supporting grants have been provided by the John Fell Fund, by the Warden and Scholars of Merton College, Oxford, and by the Craven Committee of the Faculty of Classics, University of Oxford.

<sup>5</sup> (a) was undertaken by Dr James Cummings (then Senior Academic Research Technology Specialist, IT Services, University of Oxford); (b) has been undertaken by James Chartrand, Open Sky Solutions.

<sup>6</sup> [<https://pleiades.stoa.org/>]; [<http://www.geonames.org/>]; [<https://www.eagle-network.eu/resources/vocabularies/>].

<sup>7</sup> [<https://www.zotero.org/groups/382445/isicily/items>].

<sup>8</sup> The museum database is an ongoing project initiated by Dr Michael Metcalfe, with an online interface at: [<http://sicily.classics.ox.ac.uk/museums>].

these are already published as RDF by Zotero), and during the process of conversion reference to both was incorporated on the <repository> and <bibl> elements in the TEI in anticipation of Linked Open Data. The one significant element of metadata which was normalised but not externally referenced was the dating information, and reference to, for example, [<http://perio.do/>] remains a future possibility.

The final element that was incorporated during the conversion process was the epigraphic text itself, since this was not included in the original dataset. This was done through an automated process, using available digitally published texts, exploiting the inclusion of existing digital identifiers in the original dataset (I.Sicily URIs are also aligned with Trismegistos text numbers, which facilitates further alignment with other digital epigraphic databases and corpora).<sup>9</sup> The vast majority of these texts (generously made available, e.g., by the EDR project) were themselves not originally created in EpiDoc, and so automated conversions were applied, either by providers at source (as in the case of EDR) or at the point of capture and incorporation.<sup>10</sup> Such automated transforms are not perfect, and commonly the underlying published source of the text is not captured through this process. Consequently, while more or less functional texts have been incorporated into approximately two-thirds of the EpiDoc files, all of these require human checking, further editing and appropriate attribution (all I.Sicily records carry a visible “status” indicator of “edited”, “draft” or “unchecked”). This is a pressing need, not least to ensure user-acceptance of the corpus, and is independent of the long-term aim to conduct autopsy and revision for all the inscriptions in the corpus (although the two steps can obviously be combined). At the same time, some hundreds of files remain without any data in the text division, and almost all require the inclusion of a translation. This creates both a challenge and an opportunity, which we discuss below.

The conversion was a one-time process, and subsequent editing has been managed through the use of XML editors and the interface provided by the I.Sicily website and eXist. The correspondence lists created for the upgrading of the data during conversion continue to be maintained, serving as local authority lists, in order to facilitate standardisation and external referencing in the continued editing of existing XML records and in the creation of new records. Where necessary, additional local authority lists will be created (e.g. for names and persons), when the current state of external authorities is insufficient. At present, for the purposes of data management and version control, the XML files and correspondence/authority lists are managed in an open-access GitHub repository.<sup>11</sup> For the purposes of actual digital publication and searching, the latest version of the XML records are held on a server hosted by

---

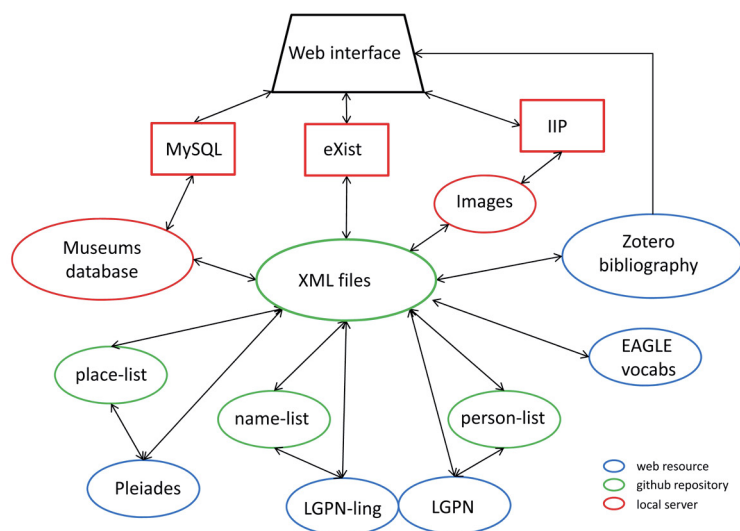
<sup>9</sup> [<http://www.trismegistos.org/>]; 2563 records currently aligned.

<sup>10</sup> [<http://www.edr-edr.it/default/index.php>]: import of converted EDR texts was kindly facilitated by Pietro Liuzzo.

<sup>11</sup> [<https://github.com/JonPrag/ISicily>].

the Faculty of Classics (University of Oxford), in an eXist database for xQuery access. URIs are maintained for the inscriptions and the museums with an eye to Linked Open Data, and both are manipulated through a RESTful API; the bibliography is published as Linked Open Data and edited directly in Zotero. The records are queried and viewed through a web interface built with AngularJS and jQuery JavaScript components. Mapping is provided in the browser by the Google Maps API. The search interface as a whole has been built very much with the difficulties of researchers in mind, exploiting new JavaScript libraries to create a spreadsheet-like interface that is flexible and reasonably intuitive, and facilitates easy export of search results.<sup>12</sup>

Images were not part of the original dataset (for the same reasons that texts were not). In the conversion, a standard template for the <facsimile> element was created in the EpiDoc, but individual image data needs to be edited into the XML files as the images become available. Currently this is a slow, manual task. We aim to make high-resolution imagery available wherever possible. In the web-interface, ZPR (Zoom, Pan, Rotate) image-viewing is provided by the IIP image server (which also enables the generation of IIIF metadata) and the OpenSeadragon JavaScript library.



**Figure 19.1:** Graphic representation of the data organisation of I.Sicily

<sup>12</sup> We gratefully acknowledge the generosity of [<https://www.ag-grid.com/>] who made the Enterprise version of their JavaScript grid available to us for free. For an overview of the search functionality of I.Sicily, see [<https://isicily.wordpress.com/how-to/>].



All of the above creates a rather complex and atomised data management structure, with XML files, authority lists, Zotero bibliography, images, and museums database held in diverse locations and curated in different ways (see Figure 19.1 for a graph). At the same time, it can be argued that this creates a very flexible system, exploiting open-source tools where possible and using standardised formats to ensure maximum interoperability, with the result that preservation and maintenance overheads are kept to a minimum. This approach is particularly well adapted to the very fluid data flows involved in curating and publishing a complex set of data that is subject to continual revision and improvement, and a continuous drip of minor updates, rather than the one-off presentation of a static dataset.

## 19.2 Challenges and Ambitions

### 19.2.1 Text-Editing and Annotation

As already noted, one of the immediate challenges faced by I.Sicily is the need to edit the text division for a large number of epigraphic texts. This task has several aspects and phases to it, each of which offers different challenges and potential solutions:

- (a) the editing of existing or missing texts, based upon published editions;
- (b) the inclusion or revision of texts based upon autopsys;
- (c) the development of a full critical apparatus for a complete edition combining (a) and (b);
- (d) the extension of mark-up, such as to record onomastic, prosopographic, or linguistic information.

With over 3,000 records, notwithstanding the fact that many are short funerary texts, this is a substantial task requiring a considerable investment of time.

Basic revision and editing (i.e. task (a)) provides a ready opportunity for developing EpiDoc training, since the I.Sicily records offer a rich set of material for students to practise editing using common tools such as the oXygen XML editor, as well as to become familiar with the basics of GitHub, which provides a convenient data management tool. At the same time, students can gain credit for their work since I.Sicily makes full use of the <resp> and <change> elements, and publishes that information in the HTML and PDF editions generated from the EpiDoc. A teaching support grant from the University of Oxford in 2015 facilitated the embedding of EpiDoc teaching within existing epigraphic teaching at the masters level, creating the necessary supply and demand relationship; and volunteer encoders have been

forthcoming.<sup>13</sup> Needless to say, such a collaborative approach requires that the documentation of the precise structure of the EpiDoc mark-up employed needs to be rigorous and available in advance in order to minimise irregularities in the edited files. The greatest challenge, however, is simply one of human resource: the resulting rapid increase in the generation of revised files, which require management and curation prior to release, creates a potential bottleneck, unless additional resources of time (or money to buy additional support) become available.

The contribution of more comprehensive revision (i.e. task (b)) based upon new information (especially through autopsy), or else of a new record for a text not previously included, in both cases including new or revised metadata, is a more challenging scenario. In principle, this can be managed through the same set of mechanisms as task (a). However, on the one hand, the free marking-up of metadata creates greater risks of irregularities; and on the other, many of those submitting such information will come from outside the academy and/or will neither have access to nor familiarity with, e.g., XML-editing (we return below to the collaborative approach responsible for this situation). Such a situation creates a need for alternative solutions to data entry, since it is both more empowering for the contributor, and more efficient for the editor, if this process can be as direct as possible (notwithstanding that a more basic approach is always possible, with an editor taking on the task of transforming data submitted in any form into a compliant XML file). At present we are experimenting with the use of an online web form,<sup>14</sup> which allows submission of a flexible range of data, while also constraining data formats for some fields and offering pre-set choices for metadata fields where authority lists exist. The form is used to generate a pre-populated XML file from the project's EpiDoc template, which is then submitted for editing. The form is still in development and, in line with the overall initial focus of the project, is again focused more on rich metadata than text-editing. A robust, web-based GUI for direct editing and revision of the actual epigraphic text remains a desideratum, but is not an immediate priority (contributors are currently left free to submit the text itself in whatever format they feel most comfortable). Pilot contributions of several sorts are underway using the form, with one set of collaborators repurposing the HTML form for use by students at a local school in Sicily (see below).

The creation of a comprehensive critical apparatus to support a final edited text (i.e. task (c) above) is a long-term desideratum, enabling the effective capture and comparison of the full information from past editions as well as fresh autopsy, but it is also a more complex challenge. In the first place, this remains an area slated for future development within the wider EpiDoc community and a relatively underdeveloped

---

<sup>13</sup> "Digital Techniques in the study of Ancient Epigraphy: transforming MSt/MPhil teaching", £ 2884 from the Humanities Division of the University of Oxford, academic year 2014/15.

<sup>14</sup> [<http://sicily.classics.ox.ac.uk/login/inscription-submit/>]. Use of the form requires the user to register, but is not restricted.

area among the majority of existing projects.<sup>15</sup> In the second place, even with such structural choices resolved, a tool that would enable editing of this part of the text mark-up, without the user having to engage directly with the increasingly complex XML involved, would be non-trivial to construct. However, examples do already exist within the wider TEI community of manuscript studies (e.g. Burghart, 2016). Part of the problem is that the demand for such an interface is more limited, since the level of already specialist knowledge entailed makes the user-group for such a tool too small to warrant the investment, at least at the scale of a project like I.Sicily. All of this implies that, in the short-term at least, this area is likely to be a significant roadblock in the final editorial development of the dataset.

A final area of text annotation, which we are currently attempting to address, is the indexing of terms within the ancient text (task (d) above). Here too, our interest lies in trying to facilitate multiple contributors, often without the ability to work directly in the XML, and not simply in resolving the problems of choosing between internal and external authority lists (where the latter even exist; see below). The two issues are, however, inter-related, since incorporating the direct referencing of external authority lists requires a different set of tools from simply building an internal list. Emblematic is the particular challenge presented by the indexing of names and individuals.<sup>16</sup> For the present, we treat the annotation of names and individuals as a discrete task, separate from general text-editing, and we are therefore content to employ a separate editing tool in order to enable the rapid annotation of names and persons across the full set of texts, by multiple contributors. The “micro-editor” for this purpose is being developed through the participation by I.Sicily in the CANARIE-funded Canadian Writers Research Collaboratory project, as one of a number of open-source tools for TEI-based projects.<sup>17</sup> We are attempting to leverage this development work with a grant from the John Fell Fund of the University of Oxford, which will permit the necessary development work within the *Lexicon of Greek Personal Names* (individuals) and the new *LGNP-Ling* database (names).<sup>18</sup> The latter will enable the publication of URIs for both named individuals in ancient Greek (i.e. persons) and names as linguistic entities, addressable via an API.

---

<sup>15</sup> Cf. [<http://www.stoa.org/epidoc/gl/latest/supp-apparatus.html>] and recent discussion on the Mark-Up list at [<http://lsv.uky.edu/scripts/wa.exe?A1=ind1710&L=markup#3>].

<sup>16</sup> See especially [<https://snapdrgn.net/>].

<sup>17</sup> Canadian Writing Research Collaboratory Extension, grant of \$CDN 456,139 for 2017-2019 from CANARIE: see [<http://beta.cwrc.ca/>].

<sup>18</sup> See [<http://www.lgpn.ox.ac.uk/>] and [<http://clas-lgpn2.classics.ox.ac.uk/>] for the XML database and [<http://admin.exist-db.org:41233/exist/apps/lgpn-ling/about.html>] for the new linguistic database of names.

### 19.2.2 Linked Open Data?

Referencing external authority lists provides an opportunity to enable greater interoperability and the creation of Linked Open Data. As has previously been observed, while EpiDoc is a huge step forward in our ability to record and represent ancient inscriptions in a rich, machine-readable, digital format, nonetheless it risks perpetuating some of the traditional challenges posed by rich but ultimately non-standardised datasets, since almost every EpiDoc project develops its own customisations and an EpiDoc file, “consists in a monolithic, self-descriptive and self-standing information unit” (Casarosa et al., 2014, p. 24, p. 28). One (partial) solution to this challenge is the use of externally referenceable controlled vocabularies – as noted above, extensive use of such reference has been incorporated into the I.Sicily EpiDoc files.

The epigraphic community has been among the leaders in the move towards the Linked Open Data approach in ancient world studies (Geser, 2016, p. 10). The stand-out example is the work of the EAGLE project, creating a set of SKOS vocabularies to enable cross-lingual referencing of core epigraphic metadata concepts.<sup>19</sup> However, as yet, the overall ontological framework has not been established to enable the full publication of EpiDoc files as RDF, and only very partial examples of the possibilities exist.<sup>20</sup> A number of reasons can be suggested (Geser, 2016 offers a thoughtful analysis in the context of archaeological data), and two might be singled out. The first, is the fact that both controlled vocabularies and referenceable authorities for many epigraphic elements are still lacking. The EAGLE vocabularies themselves are still a work-in-progress, currently lacking a clear framework for community development (this is said to be in hand), and they are not consistently adopted since they are themselves mostly aligned to larger vocabularies (e.g. DAI and Getty). As the EAGLE project itself disarmingly observes on the vocabularies landing page, “perhaps one day we will be able to do nice things as those Pelagios, Pleiades and SNAP-DRGN do [*sic*], also based on these vocabularies.” However, even the reference to SNAP-DRGN is optimistic, since currently online prosopographies themselves are a work-in-progress (the projected work on the LGPN database, referenced above, will hopefully help move this forward). The principal area where such referencing is currently possible is in the realm of geographical data. Having referenced place-name information in I.Sicily with Pleiades URIs, we have been able to generate the necessary RDF export for Pelagios, in a working demonstration of the possibilities of Linked Open Data.<sup>21</sup> However, it remains the case that for most such projects, this is currently the one effective area

<sup>19</sup> [<https://www.eagle-network.eu/resources/vocabularies/>].

<sup>20</sup> Contrast the work by the numismatic community at [<http://nomisma.org/>].

<sup>21</sup> See [<http://peripleo.pelagios.org/about>] and [<http://peripleo.pelagios.org/ui#selected=http://sicily.classics.ox.ac.uk/pelagios-data/isicily-pelagios-dataset>].

where Linked Open Data is a practical reality, and this is due to the success of the Pleiades gazetteer.<sup>22</sup> The second reason is the outstanding need to create a map from EpiDoc to a set of RDF ontologies (which entails choosing the ontologies themselves, the appropriate terms within the ontologies and, where no appropriate ontologies exist, creating a new ontology with new terms). Initial work has been undertaken on mapping EpiDoc to CIDOC-CRM (Casarosa et al., 2014) and a further discussion of epigraphic ontologies took place at the recent Open Epigraphic Data Unconference (London, 15 May 2017).<sup>23</sup> It is clear that trying to coordinate this work with others would be best in the long term, but it remains difficult to coordinate in the short term. Consequently, it remains tempting to move ahead independently and seek to publish a smaller subset of some basic RDF (as with the geographical data), mapping independently without consultation, on the assumption that such mappings could later be changed, and with the aim of encouraging further development.

In any event, I.Sicily has chosen to privilege external authority lists wherever possible, in anticipation of Linked Open Data. However, in many cases the incomplete nature of such lists means that an internal authority list is also necessary, and unless those internal lists are also maintained, published, and potentially externally aligned in the future, Linked Open Data remains a hope rather than a reality. Currently, we appear to be in something of a vicious circle, since the resource required to get Linked-Open-Data-ready is not negligible, while the demonstrable short-term (and even medium-term) gains from such activity are few and far between, meaning that there is little incentive.

### 19.2.3 Collaboration and Outreach

Although the core data of the initial instantiation of I.Sicily is derived from existing publications, moving forward we aim fully to revise each inscription record on the basis of identification of the original object and full autopsy. Such an approach is impossible without the collaboration of the museums that hold the majority of the material.<sup>24</sup> I.Sicily has therefore been constructed in a deliberately museum-centric fashion, publishing a gazetteer of Sicilian museums.<sup>25</sup> This enables the direct linking of epigraphic records to museum collections, and in turn the effective online publication of individual catalogues of museums' epigraphic collections. On the one

---

<sup>22</sup> [<https://pleiades.stoa.org/>]; [<http://commons.pelagios.org/>].

<sup>23</sup> [<https://github.com/EpiDoc/OEDUC/wiki>].

<sup>24</sup> More broadly, the overall objective of I.Sicily as a comprehensive corpus for Sicily is impossible without extensive collaboration from a wide group of experts: something which the model of digital publication with explicit attribution of responsibility clearly facilitates.

<sup>25</sup> [<http://sicily.classics.ox.ac.uk/museums>].

hand, this serves the needs of researchers who want to be able to locate individual inscriptions for study. On the other, this makes the corpus of direct value to the museums themselves, both as a service for the curatorial staff and as a potential tool for virtual display of material and other forms of increased accessibility.<sup>26</sup>

As noted above, this creates challenges in the work of collaborative recording, and we are experimenting with several models. The most productive and exciting of these to date has been a joint project with the Museo Civico Castello Ursino of Catania, the city of Catania, the Liceo artistico statale M.M. Lazzaro, and the CNR Istituto di Scienze e Tecnologie della Cognizione (ISTC) at Catania (Agodi et al., 2018). Exploiting the possibilities of the Italian Ministry of Education, Universities and Research (MIUR) “alternanza scuola-lavoro” programme (i.e. work experience for school students), we have worked with students and teachers from a large secondary school in Catania on the work of cataloguing the epigraphic collection of the Catania civic museum. A group from the CNR-ISTC (the “EpiCUM project” directed by Dr Daria Spampinato) has in turn worked with the students, developing a version of our own HTML record form to enable the students to input data into an automatically generated XML file. The CNR-ISTC project is in turn using the I.Sicily template for a digital catalogue of the non-Sicilian inscriptions in the collection (EpiCUM). All parties worked together to curate a permanent exhibition (“Voci di pietra”) in the museum of a selection of 35 inscriptions from ancient Catania, which opened on 14 July 2017.<sup>27</sup> The EpiCUM project is also developing a parallel virtual exhibition, in part based upon the I.Sicily EpiDoc files. The students undertook cleaning, recording and conservation work in the museum prior to the exhibition, and played a leading role in the design and production of the exhibition itself. Subsequently, they have continued cataloguing and recording the c. 500 inscriptions in the museum’s collection. With additional funding from the University of Oxford, a follow-up collaboration is now being planned with a second Liceo at the Museo Archeologico Regionale “Paolo Orsi”, in Siracusa. An approach of this sort creates many problems of its own, but two very clear advantages can be observed: firstly, a very much more rapid aggregation of (genuinely high quality) data; secondly, a real sense of community engagement and empowerment, bringing local epigraphic material into the public consciousness, rendering it comprehensible as ‘voices of stone’ from a community’s past.

---

<sup>26</sup> Note e.g. the *izi.travel* project, which is very active in Sicily and which can link to I.Sicily for epigraphic objects (see [<https://izi.travel/it/4d91-museo-archeologico-regionale-paolo-orisi/it>] for an example of a museum tour).

<sup>27</sup> The exhibition was supported by a Knowledge Exchange Fellowship from The Oxford Research Centre in the Humanities, 2016/17. Press coverage includes: [[http://www.corriere.it/foto-gallery/cultura/17\\_agosto\\_03/catania-romana-raccontata-voci-pietra-48d4802e-77ec-11e7-84f5-f24a994b0580.shtml](http://www.corriere.it/foto-gallery/cultura/17_agosto_03/catania-romana-raccontata-voci-pietra-48d4802e-77ec-11e7-84f5-f24a994b0580.shtml)] and [<http://www.globusmagazine.it/110708-2/#.WiR9MDdpGwV>].

### 19.3 Conclusions

There are a number of further challenges presented by the I.Sicily corpus which we have not considered here, such as the complications presented by a very non-uniform corpus covering not only a very extended period in time (and so, e.g., Archaic texts compared to Christian texts), but also an increasingly wide variety of materials, and in particular a rich mixture of languages, not all of which have a Unicode character set. From a practical perspective, the current state of the relevant technologies and limited availability of resources makes an undertaking of this sort extremely challenging, above all if one seeks to build an open, collaborative project, rather than a closed, local dataset resulting in a static publication. From a purely scientific perspective, the greatest challenge remains the acceptance not only of a born-digital publication, but also of a publication that is not stable in the traditional sense and has no clear single publication date. Transparency and rigorous, detailed attribution of responsibility appear, to us, to be the most effective responses to this, hopefully temporary, problem. Nonetheless, we have been hugely encouraged by the enthusiasm with which colleagues, museums, local authorities, and local communities have embraced the project so far, and we remain fundamentally optimistic about the potential for the future – not least because of the strength of the EpiDoc community itself.

### Bibliography

- Agodi, S, Cristofaro, S., Noto, V., Prag, J., & Spampinato, D. (2018). Una collaborazione tra museo, enti di ricerca e scuola: l'epigrafia digitale e l'alternanza scuola lavoro. *Umanistica Digitale*, 2. doi: 10.6092/issn.2532-8816/7298. Retrieved from [https://umanisticadigitale.unibo.it/article/view/7298], 2018/06/18.
- Bodard, G. (2010). EpiDoc: Epigraphic Documents in XML for Publication and Interchange. In F. Feraudi-Gruénais (Ed.), *Latin on Stone: Epigraphic Research and Electronic Archives* (pp. 1–17). Lanham, MD: Lexington Books.
- Burghart, M. (2016). The TEI Critical Apparatus Toolbox: Empowering Textual Scholars through Display, Control, and Comparison Features. *Journal of the Text Encoding Initiative*, 10. doi: 10.4000/jtei.1520. Retrieved from [http://jtei.revues.org/1520], 2017/11/17.
- Casarosa, V., Manghi, P., Mannocci, A., Rivero Ruiz, E., & Zoppi, F. (2014). A Conceptual Model for Inscriptions. In S. Orlandi, R. Santucci, V. Casarosa, & P.M. Liuzzo (Eds.), *Information Technologies for Epigraphy and Cultural Heritage. Proceedings of the First EAGLE International Conference, Paris* (pp. 23–40). Rome: Sapienza Università Editrice. Retrieved from [https://www.eagle-network.eu/wp-content/uploads/2015/01/Paris-Conference-Proceedings.pdf], 2017/12/03.
- De Vido, S. (1999). Corpora epigrafici siciliani da Gualtherus a Kaibel. In M.I. Gulletta (Ed.), *Sicilia Epigraphica. Atti del convegno internazionale, Erice, 15-18 ottobre 1998* (vol. I, pp. 221–250). Pisa: Edizioni della Normale.
- Fazello, T. (1558). *F. Thomæ Fazelli... de rebus Siculis decades duæ*. Panormi: Panormi.
- Geser, G. (2016). *Towards a Web of Archaeological Linked Open Data* (ARIADNE WP15 Study. V1.0, 6 October 2016). Salzburg Research, Austria. Retrieved from [http://www.

- ariadne-infrastructure.eu/content/download/8392/49194/version/2/file/ARIADNE\_archaeological\_LOD\_study\_10-2016.pdf], 2017/12/03.
- Gualtherus, G. (1624). *Siciliæ obiacentium insular. et Bruttiorum antiquæ tabulæ, cum animadversionib.* Messanae: apud Petrus Bream.
- Gulletta, M.I. (Ed.). (1999). *Sicilia Epigraphica. Atti del convegno internazionale, Erice, 15-18 ottobre 1998* (2 vols.). Pisa: Edizioni della Normale.
- Kaibel, G. (1890). *Inscriptiones Italiae et Siciliae* (= IG XIV). Berlin: Georgius Reimerus.
- Manganaro, G. (1988). La Sicilia da Sesto Pompeo a Diocleziano. *Aufstieg und Niedergang der römischen Welt*, 2.11.1, 3–89.
- Mommsen, T. (1883). *Inscriptiones Bruttiorum Lucaniae Campaniae Siciliae Sardiniae Latinae. Pars posterior. Inscriptiones Siciliae et Sardiniae* (= CIL X.2). Berlin: G. Reimer.
- Prag, J.R.W. (2002). Epigraphy by numbers: Latin and the epigraphic culture in Sicily. In A.E. Cooley (Ed.), *Becoming Roman, Writing Latin?* (Journal of Roman Archaeology Supplementary Series 48) (pp. 15–31). Portsmouth, RI.
- Prag, J.R.W. (2003). Nouveaux regards sur les élites locales de la Sicile républicaine. *Histoires et sociétés rurales*, 19, 121–132.
- Prag, J.R.W. (2007). Ciceronian Sicily: the epigraphic dimension. In J. Dubouloz & S. Pittia (Eds.), *La Sicile de Cicéron, Lectures des Verrines* (pp. 245–271). Besançon: Presses universitaires de Franche-Comté.
- Prag, J.R.W. (2008). Sicilia and Britiannia: Epigraphic Evidence for Civic Administration. In C. Berrendonner, M. Cébeillac Gervasoni, & L. Lamoine (Eds.), *Le Quotidien municipal dans l'Occident romain* (pp. 67–81). Clermont-Ferrand: Presses Universitaires Blaise-Pascal.
- Prag, J.R.W. (2010). Sicilia Romana tributim discripta. In M. Silvestrini (Ed.), *Le tribù romane. Atti della XVII Rencontre sur l'épigraphie (Bari 8-10 ottobre 2009)* (pp. 305–311). Bari: Edipuglia.
- Tribulato, O. (Ed.). (2012a). *Language and linguistic contact in ancient Sicily*. Cambridge: Cambridge University Press.
- Tribulato, O. (2012b). *Siculi bilingues?* Latin in the inscriptions of early Roman Sicily. In Tribulato, O (Ed.), *Language and linguistic contact in ancient Sicily* (pp. 291–325). Cambridge: Cambridge University Press.



## Conclusions

The contributions collected in this volume, in particular those regarding the “marginal” epigraphies, bring to attention a considerable variety of themes, even beyond those initially envisaged while conceiving the volume. Even though a systematic vision and approach in digital epigraphy is still very far off, a summary of the different issues and positions highlights the common trends.

Albeit text and text-bearing object cannot be separated in the study of epigraphy, it is apparent that the text is the main focus in digitization. The recurring objective of epigraphic projects is making strings of inscriptional characters searchable, which may or may not include lacunae, integrations, variants and corrections. Presently, the most common practice is the transcription/transliteration and the XML encoding of different kinds of phenomena (structural, concerning the relationship with support, textual portions, transcription phenomena, editorial interventions, in-line *apparatus criticus*, PoS and morphological analysis, onomastics, etc.), depending on the topics and objectives of the projects. EpiDoc, a subset of elements of the TEI standard, is widely used both as archiving and exchange format. The agreement of the scientific community regarding well-defined best practices is an important achievement. Nevertheless, this method is not suitable for all the epigraphic materials, depending on the writing systems used and on the current degree of knowledge of the scripts and languages attested.

The contributions on the Maya, Linear B and cuneiform scripts witness alternative solutions with respect to the XML encoding of the texts, in relation to logo-syllabic writing systems. The *Sinleqiunnini* project (Di Filippo) shows the application of the relational model to the texts themselves, according to the “ordered hierarchy of content objects” theoretical framework. The limits of XML encoding in the annotations of not-contiguous portions of text, and in the management of overlapping hierarchies, are thus overcome. As texts are the sum of the instances of hierarchical entities, conflicting interpretations of the phonetic, morphological and semantic values can be recorded.

The *Text Database and Dictionary of Classic Mayan* (Prager et al.), on the other hand, is implementing a sign catalogue, in addition to the corpus of TEI encoded texts, which identifies each graph and relates it to its allographs using a propositional logic. The semantic modelling (CIDOC CMR and RDF encoding) allows for different, duly argued, readings.

The *RuneS* project (Zimmermann, Kezzazi, & Bahr) has a further approach. The different systems of Runic script are the focus of the project: the philological study of the graphic variants of signs is carried out through the annotation of the visual documentation, photographs in particular, which leads to the creation of a catalogue of signs. Such an approach, even though it does not allow the textual search, broadens the possibility of the palaeographic research, and is currently supported by the spread of standards for the interoperability of images, such as IIIF.

Even though the Palaeohispanic inscriptions of the *Hesperia* project (Estarán et al.) are transliterated in Latin characters, several phonetic interpretations, deriving from different editions of the epigraph, can be attributed to each graph so that the search engine retrieves all variants.

Those experiences attest to the diverse approaches to the structuring and archiving of textual data, depending on the characteristics of the epigraphic material and the objectives of the digitization project.

The practice of storing textual and extra-textual data in relational databases is indeed more widespread than the establishment of repositories of XML files for the same purpose. This is the case even for collections of texts written in fully deciphered scripts, whose transliteration and interpretation is plain. The modelling of discrete entities entails the independent descriptions of the various aspects of the inscriptional document, leading to the enrichment of information on the physical carriers – thus stimulating studies in the fields of material culture, iconography and history of art – or on their places of origin and provenance (Xella & Zamora). The implementation of contextual information, as discussed further on, is the foundation for establishing connections with archaeological datasets, and developing tools that support interoperability, such as chronological and geographical gazetteers.

Epigraphs are often the only sources for the study of languages whose fragmentary attestation currently hinders their complete comprehension. The digitization of textual corpora provides a host of data to be processed with the support of technology. However, lexicographic developments vary considerably, depending on languages and digitizing methods. In general, an automatic approach is preferred when very large digitized corpora are available.

NLP techniques are fruitful when applied to modern, living languages, which benefit from the mass of data collected from the web, as well as to Greek and Latin among the languages of antiquity. Greek and Latin have a considerable volume of digitized texts, which also include non-epigraphic documents such as papyri and manuscripts. Moreover, the availability of reference tools such as dictionaries and grammars, allows performing the automatic parsing of texts, PoS and morphological analysis, by matching each word with a fixed set of possible lemmata and word-formation rules.

On the contrary, a “manual” approach is preferred by projects focusing on languages, which are fragmentarily attested and not provided with those tools (the so-called “under-resourced languages”), or languages whose signs’ value or signs’ sequence interpretation is multiple. Some of them encode grammatical phenomena in the texts. Morphological analysis is carried out for each lexeme, which is then connected to a lemma and, in the case of Semitic languages, to a root as exemplified by the papers on the OIMEA (Novotny & Radner), OCIANA (Burt, Al-Jallad, & Macdonald) and Sabäisches Wörterbuch (Multhoff) projects. The strong repetitiveness of texts and, at the same time, the morphological ambiguity of the languages that make the direct encoding burdensome but still uncertain, may otherwise suggest the creation

of independent, lexical entries to be linked to occurrences in the texts, as in the DASI lexicon (Avanzini, De Santis, & Rossi).

KALAM (Ruzicka) is the first attempt at conducting an automatic detection of morphological attributes in pre-Islamic Arabian texts, on the basis of the generative grammar and with a synthetical approach. In any case, the lexical tools that are being developed will increasingly improve the linguistic knowledge and the available dataset, thereby enabling an automatic approach for these languages.

Given the specific needs and the differences so far described, we must underline a patent, yet still somehow underestimated issue: the boundaries of disciplines and methods should not affect the study of the societies of the past. A thorough knowledge is reached via a comprehensive approach, in which interoperability and open access to data play an essential role.

The contribution by Liuzzo clearly exemplifies the challenges to be faced when aggregating a huge amount of epigraphic records, although quite homogenous from the historical and the linguistic points of view. A complex harmonization of the tributary vocabularies has been necessary in the EAGLE project. If EpiDoc remains the most suitable schema for the interchanging of epigraphic records (also by virtue of its mapping to RDF), the exploitation of established thesauri and ontologies is a key to fostering interoperability – as exemplified by the case of the research platform EPIDAT for Jewish epigraphy (Kollatz). These tools force us to face the issue of matching the taxonomies that make sense in each domain. As for the encoding aspect, even if the EAGLE schema is based on the EpiDoc standard, the automatic up-conversion of the records provided by partners, has sometimes produced unsatisfactory results due to the personal and varied use of the EpiDoc elements by different providers.

It is right to ask ourselves whether the best practice is to conform to strict common operational guidelines, or to maintain diversity within a shared general framework, enacting mapping strategies when needed. In fact, the discrepancy in the choice of the phenomena to annotate, and the encoding solutions (though in the frame of the same guidelines), are usually objective-driven and therefore answer to different scientific questions. This is especially apparent when dealing with several editions for the same text across distinct repositories.

To preserve this wealth of information, and at the same time offering a single point of access to all the digital editions of the texts while disambiguating them, comprehensive indexes of texts – harvesting the existing repositories – are desirable. The project Trismegistos (Depaw), moves towards this objective by providing a unique ID for each textual document, and is now expanding its scope beyond the indexing of texts to the indexing of people and places mentioned in those sources, from a LOD perspective.

Encoding of onomastics is carried out by almost all of the projects represented in this volume, in some projects in addition to named entity recognition. The first successful attempts at connecting information and producing historical knowledge, through prosopographical authority files and geographical gazetteers, are very

promising. The identification of persons mentioned in the texts, the reconstruction of lineages, and the network analysis will enhance cross-referencing.

The general cooperation with initiatives like Pelagios and PeriodO (see Rabinowitz, Shaw, & Golden), and the references to their gazetteers of places and periods, would reduce the obstacles posed by the naming of geographical entities, which are continuously changing in relation to the historical context, and the dating systems used by different socio-political entities in the sources, as well as by different traditions of study.

As regards the interoperability of multilingual textual corpora, epigraphs witness, perhaps better than other sources, the cohabitation of different languages and scripts in the same periods and in the same regions (see the example of the same inscription attested in different languages in Bausi & Liuzzo), or the use of a language or script in very distant geographic, socio-political or historical contexts. Is it possible and worthwhile to imagine a simultaneous search on texts in different languages, and on concepts, across different corpora? The exploitation of lexical semantics and translations deserves deeper attention by the community of epigraphers in order to enjoy their full potential in terms of linguistic and cross-cultural research. Translations are not envisaged in many digitization projects. Indeed, their recording requires us to choose whether to provide a literal translation, or to include periphrases and metaphors, and then require conformity to this choice across the overall corpus. The standardization of translations within each project should be granted, and a strict relation between segments of texts and their translations should be envisaged, if not a proper encoding of the translations. Moreover, attempts at identifying and mapping common concepts in different corpora, at least those linguistically and culturally close, might produce unexpected results.

Operating at a translation level would allow scholars not familiar with a particular language to consider sources potentially useful for historical and cultural studies. Furthermore, as well as valorising the physicality of the inscriptions and their iconographic apparatus, translations would also make epigraphy more interesting for, and accessible by, a wider public. Projects as *I.Sicily* (Prag & Chartrand) are stimulating sensitivity towards the preservation and appreciation of the peculiar epigraphic historical witness, by operating within local cultural or educational institutions and by envisaging digital tools for the cross-cultural fruition of the sources within a defined geographic context.

The commitment to the digital preservation of the epigraphic heritage through a thorough, systematic collection of photographs is at the centre of many projects, acknowledging the importance of autopsy in the study of primary material sources, as exemplified by the *Karnak* project (Biston-Moulin & Thiers). Free access to and reuse of images becomes, thus, a key issue for the progress in epigraphic research, feeding the ongoing debate on the need of softening – if not removing – copyright obstacles by museums, archives, libraries, and even archaeological sites. This is

especially needed in contexts at risk of deterioration or destruction by human and/or environmental factors.

Last but not least, one major challenge emerging from the contributions in this volume is the sustainability of projects and their results.

This is probably one of the main factors that most discourages the creation of critical editions in digital format and the commitment to their implementation. Many scholars continue to prefer paper editions, as the editorial criteria of many projects do not allow the attribution of contributions and interventions to each individual researcher, thus impeding the proper citation of digital resources, and therefore the evaluation of his/her scientific activity. Moreover, traditional publications are still the preferred medium for a detailed communication to, and easier fruition by, the scientific community of the research's results (as stated by Cannata, and Biston-Moulin & Thiers), thus receiving a better appreciation by the specialized audience.

Furthermore the general fragility and volatility of digital knowledge threatens the total disappearance of scholars' scientific work in the medium term. The earliest, pioneering projects have not always had the opportunity to upgrade their technical infrastructure. The rapid, technical obsolescence, despite the rigorous scientific methodology, endangers access to their data.

Sooner or later, every digital epigraphy "venture" has to face this problem, due to the lack of permanent funding. Some projects are included in larger portals, like the one hosting and supporting Hittitology projects described in the paper by Müller & Schwemer. Some take the legal form of associations and foundations, and benefit from the support provided by their infrastructures: the IDEA initiative described by Liuzzo is the first attempt at exploring this path in the digital epigraphy domain. More often, engaging with new initiatives is the only viable solution (when achieved) to maintain previous (well established, but no more attractive) ones. Openness and reuse is the suggested best practice to circulate, multiply and save, at least, data.

Indeed, the definition of a common, theoretical and methodological framework that will make this research effective is only possible if we take into account the variety of questions, problems, approaches and solutions dealt with by the widest community of epigraphers. This volume was an attempt at gathering, virtually, at a "round table" some of those who practice digital epigraphy in very different cultural domains and with different scopes, and who have been engaged in this kind of research activity over different periods of time. They have shared their experiences to stimulate discussion of some of the main themes that are driving digital epigraphy forward in the future, while also shaping our approach to traditional epigraphy.

# Appendix A

## Selected Webliography

This webliography includes a selection of the online resources that have been referenced within the papers. Among them, only those useful to approach digital epigraphy, in content and method, have been selected. Each resource is described through the core elements of the Dublin Core Metadata Initiative. Therefore, especially indications on subjects, and chronological and geographic coverage are general, not domain-specific.

## Archives of Digitized Inscriptions and Aggregators

1.

Title: Ancient European Languages and Writings – AELAW

Description: AELAW is dedicated to the study of the different ancient European languages and writings with the objective of a large online databank which will permit the cataloguing of all the currently known documents in this type of languages, thus introducing this important part of the European cultural heritage into the 21st century.

Identifier: [<http://aelaw.unizar.es/>]

Creator: AELAW

Date: 2018

Format: text/html

Language: eng

Subject: Inscriptions [<http://id.loc.gov/authorities/subjects/sh85066566>]; Extinct languages [<http://id.loc.gov/authorities/subjects/sh85046567>]

Coverage: Europe [<http://vocab.getty.edu/tgn/1000003>]; Antiquité [<http://n2t.net/ark:/99152/p0qhb66qj4c>]; Iron Age [<http://n2t.net/ark:/99152/p0ff3dt8qvz>]

Type: Text, Dataset

2.

Title: The Ancient Graffiti Project – AGP

Description: The Ancient Graffiti Project focuses on handwritten inscriptions of the early Roman empire, especially in Herculaneum and Pompeii. The aim of AGP is to allow scholars and the public to explore ancient handwritten wall-inscriptions and to understand them in context. It provides maps to help viewers understand where graffiti appeared in the ancient city and offers translations and brief summaries of the graffiti. The inscriptions presented are critical editions of the ancient texts, many of which offer updates to the Corpus Inscriptionum Latinarum.

Identifier: [<http://ancientgraffiti.org/Graffiti/>]

Creator: Washington and Lee University

Date: 2018

Format: text/html

Language: eng

Subject: graffiti [<http://id.loc.gov/vocabulary/ethnographicTerms/afset008057>];  
Inscriptions, Greek [<http://id.loc.gov/authorities/subjects/sh85066590>];  
Inscriptions, Latin [<http://id.loc.gov/authorities/subjects/sh85066606>]

Coverage: Pompeii (deserted settlement) [<http://vocab.getty.edu/tgn/7004658>];  
Herculaneum (deserted settlement) [<http://vocab.getty.edu/tgn/7031897>]

Type: Dataset

3.

Title: Archives babyloniennes (XXe–XVIIe siècles av. J.-C.) – ARCHIBAB

Identifier: [[www.archibab.fr](http://www.archibab.fr)]

Description: The ARCHIBAB project has among its objectives the creation of a database containing the edition of all the archive documents (letters, legal texts and economic texts) dated to the Old Babylonian period.

Creator: Collège de France - Dominique Charpin, Antoine Jacquet

Date: 2008

Format: text/html

Language: fra

Subject: Assyro-Babylonian [<http://id.loc.gov/authorities/subjects/sh85008838>];  
Akkadian language—Texts [<http://id.loc.gov/authorities/subjects/sh2007100963>]

Coverage: Mesopotamia (general region) [<http://vocab.getty.edu/tgn/7001554>]; Old  
Babylonian/Assyrian Mesopotamia (2000–1600 BC)

Type: Dataset

4.

Title: Beta maṣāḥəft: Manuscripts of Ethiopia and Eritrea

Description: The project aims at creating a research environment that shall manage complex data related to the Christian manuscript tradition of the Ethiopian and Eritrean Highlands. Manuscript descriptions, accompanied by images, shall be made available and searchable, and various texts shall be edited. In addition, a comprehensive prosopography and a historical gazetteer of Christian Ethiopian culture shall emerge, alongside a digital Clavis of literature in Ethiopic.

Identifier: [<https://www.betamasaheft.uni-hamburg.de/>]

Creator: Universität Hamburg

Date: 2018

Format: text/html

Language: eng

Subject: Manuscripts, Ethiopic [<http://id.loc.gov/authorities/subjects/sh85080711>];  
Inscriptions, Ethiopic [<http://id.loc.gov/authorities/subjects/sh94000935>]

Coverage: Ethiopia (nation) [<http://vocab.getty.edu/tgn/7000489>]; Eritrea (nation) [<http://vocab.getty.edu/tgn/7001658>]; Aksumite [<http://n2t.net/ark:/99152/p03tcss4qv>]

Type: Text; Dataset

5.

Title: Cachette de Karnak

Identifier: [<http://www.ifao.egnet.net/bases/cachette/>]

Description: The Cachette de Karnak project is concerned with the online publication of the materials coming from north-west of the courtyard of the 7th pylon in the Temple of Karnak, where over 700 statues in stone, 17000 in bronze and many other artefacts were discovered in 1903 by the archaeologist G. Legrain. The database gives for each object a general description, photographic documentation, its different registration numbers and a bibliography.

Creator: Institut français d'archéologie orientale

Date: 2017

Format: text/html

Language: fra, eng

Subject: Inscriptions, Egyptian [<http://id.loc.gov/authorities/subjects/sh85041341>]

Coverage: Egypt (former nation/state/empire) [<http://vocab.getty.edu/tgn/7014986>]; Karnak (deserted settlement) [<http://vocab.getty.edu/page/tgn/7764757>]; Old Kingdom Egypt (2670–2168 BCE/BC) [<http://n2t.net/ark:/99152/p03wskdgmftf>]; First Intermediate Period Egypt (2168–2010 BCE/BC) [<http://n2t.net/ark:/99152/p03wskdxjnj>]; Middle Kingdom Egypt (2010–1640 BCE/BC) [<http://n2t.net/ark:/99152/p03wskdq4n>]; Second Intermediate Period Egypt (1640–1548) [<http://n2t.net/ark:/99152/p03wskdzd99>]; New Kingdom Egypt (1548–1086) [<http://n2t.net/ark:/99152/p03wskddb3j>]; Third Intermediate Period Egypt (1086–664) [<http://n2t.net/ark:/99152/p03wskdmzfr>]; Late Period Egypt (664–332) [<http://n2t.net/ark:/99152/p03wskd47fw>]; Macedonian Egypt (332–304 BCE/BC) [<http://n2t.net/ark:/99152/p03wskdxnwr>]; Ptolemaic-Roman Egypt (304 BC–AD 640) [<http://n2t.net/ark:/99152/p03wskdftkm>]

Type: Dataset

6.

Title: Celtic Inscribed Stone Project – CISP

Description: CISP aims at a collaborative, interdisciplinary study of Medieval Celtic inscriptions. One of its main objectives is the compilation of an accessible, comprehensive and authoritative database of all known inscriptions, including those brought to light in the field work undertaken in Brittany and the Channel Islands. The scope of the project is the Celtic-speaking regions of the early middle ages, (Scotland, Ireland, Wales, Brittany, the Isle of Man, and parts of western England, in the period approximately AD 400–1100). Included are all stone monuments inscribed with text,



whether in the Celtic vernacular or Latin, in the Roman alphabet or Ogham (but excluding runic inscriptions).

Identifier: [<http://www.ucl.ac.uk/archaeology/cisp>]

Creator: University College London

Date: 1999

Format: text/html

Language: eng

Subject: Inscriptions, Celtic [<http://id.loc.gov/authorities/subjects/sh97000858>]; Celtic languages [<http://id.loc.gov/authorities/subjects/sh85021721>]; Latin language [<http://id.loc.gov/authorities/subjects/sh85074944>]; Ogham alphabet [<http://id.loc.gov/authorities/subjects/sh85094245>]

Coverage: Brittany (historical region) [<http://vocab.getty.edu/tgn/7024267>]; Scotland (country) [<http://vocab.getty.edu/tgn/7002444>]; Ireland (island) [<http://vocab.getty.edu/tgn/7001181>]; Wales (country) [<http://vocab.getty.edu/tgn/7002443>]; Isle of Man (island) [<http://vocab.getty.edu/tgn/7005260>]; Early Medieval [<http://n2t.net/ark:/99152/p0kh9dsmf3f>]

Type: Dataset

7.

Title: Corpus Inscriptionum Phoenicarum necnon Poenicarum – CIP

Description: The Corpus Inscriptionum Phoenicarum necnon Poenicarum is a project which collects and produces a critical edition of all the Phoenician and Punic epigraphic documents.

Identifier: [<http://cip.cchs.csic.es/>]

Creator: Instituto di Studi sul Mediterraneo Antico - CNR; Centro de Ciencias Humanas y Sociales - CSIC, Madrid

Date: 2018

Format: text/html

Language: eng

Subject: Inscriptions, Phoenician [<http://id.loc.gov/authorities/subjects/sh85066622>]; Inscriptions, Punic [<http://id.loc.gov/authorities/subjects/sh85066626>]

Coverage: Iron Age [<http://n2t.net/ark:/99152/p0f65r2nwf7>]; Persian [<http://n2t.net/ark:/99152/p0f65r2s5gc>]; Hellenistic-Roman Early Empire (330 BC–AD 300) [<http://n2t.net/ark:/99152/p03wskd825s>]; Punic [<http://n2t.net/ark:/99152/p08m57hph3k>]

Type: Dataset

8.

Title: Cuneiform Digital Library Initiative – CDLI

Description: The Cuneiform Digital Library Initiative is an international digital library project aimed at putting text and images of an estimated 500,000 recovered cuneiform tablets created from between roughly 3350 BC and the end of the pre-Christian era online.

Identifier: [<https://cdli.ucla.edu>]

Creator: University of California; University of Oxford; Max Planck Institute for the History of Science

Date: 2018

Format: text/html

Language: eng

Subject: Cuneiform inscriptions [<http://id.loc.gov/authorities/subjects/sh85034803>]

Coverage: Mesopotamia (general region) [<http://vocab.getty.edu/tgn/7001554>]

Type: Dataset

9.

Title: Danske Runeindskrifter

Description: The database Danske Runeindskrifter is a presentation of all Danish Rune inscriptions prepared in a three-year cooperation project between the Nordic Research Institute at the University of Copenhagen and the National Museum in Copenhagen.

Identifier: [[www.runer.ku.dk](http://www.runer.ku.dk)]

Creator: University of Copenhagen; National Museum of Copenhagen

Date: 2009

Format: text/html

Language: dan

Subject: Inscriptions [<http://id.loc.gov/authorities/subjects/sh85066566>]; Runes [<http://id.loc.gov/authorities/subjects/sh85115851>]

Coverage: Northern Europe [<http://vocab.getty.edu/tgn/4003757>]

Type: Dataset

10.

Title: Database of Neo-Sumerian Texts – BDTNS

Description: The Database of Neo-Sumerian Texts (or BDTNS, its acronym in Spanish) is a searchable electronic corpus of Neo-Sumerian administrative cuneiform tablets dated to the 21st century BCE.

Identifier: [<http://sefarad.filol.csic.es>]

Creator: Consejo Superior de Investigaciones Científicas

Date: 2015

Format: text/html

Language: eng

Subject: Cuneiform inscriptions, Sumerian [<http://id.loc.gov/authorities/subjects/sh85034807>]

Coverage: Mesopotamia (general region) [<http://vocab.getty.edu/tgn/7001554>]; Neo-Sumerian [<http://n2t.net/ark:/99152/p083p5rcd2r>]

Type: Dataset

11.

Title: Deutsche Inschriften Online – DIO

Description: The goal of the project is the digitization and online provision of the inscriptions edited in the volumes of the Deutsche Inschriften series. This collects all the Latin and German inscriptions of the Middle Ages and the Early Modern period up to the year 1650. The collection area includes the current state of Germany and Austria and South Tyrol.

Identifier: [<http://www.inschriften.net>]

Creator: Akademie der Wissenschaften zu Göttingen; Akademie der Wissenschaften und der Literatur Mainz

Date: 2018

Format: text/html

Language: deu

Subject: Inscriptions, Latin [<http://id.loc.gov/authorities/subjects/sh85066606>]; Inscriptions, German [<http://id.loc.gov/authorities/subjects/sh2004005912>]Coverage: Germany (nation) [<http://vocab.getty.edu/tgn/7000084>]; Austria (nation) [<http://vocab.getty.edu/tgn/1000062>]; South Tyrol (general region) [<http://vocab.getty.edu/tgn/7030436>]; Mittelalter [<http://n2t.net/ark:/99152/p0qhb662qrr>]; Modern History, Period I [<http://n2t.net/ark:/99152/p0jf288v75n>]

Type: Dataset

12.

Title: Digital Archive for the Study of pre-Islamic Arabian Inscriptions – DASI

Description: DASI seeks to gather all known pre-Islamic Arabian epigraphic material into a comprehensive online database, with the aim to make available to specialists and to the broader public a wide array of documents often underestimated because of their difficulty of access. By means of a digitization process through a hybrid data entry/xml system according to international encoding standards, DASI gives access at present to nearly 8,000 Ancient South Arabian inscriptions recorded by the University of Pisa team.

Identifier: [<http://dasi.cnr.it/>]

Creator: University of Pisa; Scuola Normale Superiore di Pisa

Date: 2013

Format: text/html

Language: eng

Subject: Inscriptions [<http://id.loc.gov/authorities/subjects/sh85066566>]; Epigraphic South Arabian language [<http://id.loc.gov/authorities/subjects/sh98000742>]; Inscriptions, Lihyanic [<http://id.loc.gov/authorities/subjects/sh85066608>]; Inscriptions, Nabataean [<http://id.loc.gov/authorities/subjects/sh85066615>]; Arabian Peninsula–History–To 622 [<http://id.loc.gov/authorities/subjects/sh2006007375>]Coverage: Arabian Peninsula [<http://vocab.getty.edu/tgn/1012700>]; Arabian (culture) [<http://vocab.getty.edu/page/aat/300019797>]

Type: Dataset

13.

Title: Ebla Digital Archives – EbDA

Description: The Ebla Digital Archives [EbDA] aims to provide a digital edition of the entire corpus of Ebla texts. It includes all documents published so far in the ARET series (“Archivi Reali di Ebla – Testi”) as well as in other monographs and journals. The digital edition provides harmonized transliterations, corrections and numerous collations. Users may browse the documents individually, or query data in the most flexible way, thanks to one of the most advanced database implementation for the digital representation of cuneiform documents. An extensive, searchable, up-to-date bibliography of all Ebla material published so far complements the results.

Identifier: [<http://ebda.cnr.it/>]

Creator: Università Ca' Foscari Venezia; CNR-Istituto di Studi sul Mediterraneo Antico  
Date: 2018

Format: text/html

Language: eng

Subject: Cuneiform inscriptions [<http://id.loc.gov/authorities/subjects/sh85034803>];

Eblaite language—Texts [<http://id.loc.gov/authorities/subjects/sh2009124035>]

Coverage: Tell Mardikh (deserted settlement) [<http://vocab.getty.edu/tgn/7002266>];

Ancient Syria (general region) [<http://vocab.getty.edu/tgn/8711750>]; Early Bronze Age

III [<http://n2t.net/ark:/99152/p0m63njtn97>]

Type: Dataset

14.

Title: The Electronic Text Corpus of Sumerian Literature – ETCSL

Description: The Electronic Text Corpus of Sumerian Literature comprises a selection of nearly 400 literary compositions recorded on sources which come from ancient Mesopotamia and date to the late third and early second millennia BCE. The corpus contains Sumerian texts in transliteration, English prose translations and bibliographical information for each composition. The transliterations and the translations can be searched, browsed and read online using the tools of the website.

Identifier: [<http://etcsl.orinst.ox.ac.uk>]

Creator: University of Oxford

Date: 2016

Format: text/html

Language: eng

Subject: Sumerian literature [<http://id.loc.gov/authorities/subjects/sh85130415>]

Coverage: Mesopotamia (general region) [<http://vocab.getty.edu/tgn/7001554>]; Early

Dynastic Mesopotamia (2950–2350 BC) [<http://n2t.net/ark:/99152/p03wskdsdnb>];

Akkadian-Ur III Mesopotamia (2335–2000 BC) [<http://n2t.net/ark:/99152/p03wskdbvmg>];

Insin-Larsa [<http://n2t.net/ark:/99152/p047fhm6w33>]

Type: Dataset

15.

Title: Epigraphic Database Heidelberg – EDH

Description: The Epigraphic Database Heidelberg contains the texts of Latin and bilingual (i.e. Latin-Greek) inscriptions of the Roman Empire. With the help of search functions specific queries can be carried out e.g. a search for words in inscriptions and/or particular descriptive data. The search results are often displayed together with photos and drawings. The geographic focus is provided by the provinces of the Roman Empire.

Identifier: [<https://edh-www.adw.uni-heidelberg.de/>]

Creator: Heidelberg Academy of Science and Humanities

Date: 2018

Format: text/html

Language: deu; eng

Subject: Inscriptions, Latin [<http://id.loc.gov/authorities/subjects/sh85066606>];

Inscriptions, Greek [<http://id.loc.gov/authorities/subjects/sh85066590>]

Coverage: Roman Empire (former nation/state/empire) [<http://vocab.getty.edu/tgn/7030347>]

Type: Dataset

16.

Title: Epigraphic Database Rome – EDR

Description: EDR focuses on the ancient inscriptions from Rome, the Italian peninsula, Sicily and Sardinia. It carries out the registration of the Greek and Latin inscriptions, except the Christian ones, prior to the 7th century AD, according to the best existing edition, possibly with further checks and amendments and with the backing of some fundamental data and images.

Identifier: [<http://www.edr-edr.it/>]

Creator: Università La Sapienza di Roma

Date: 2018

Format: text/html

Language: ita; eng

Subject: Inscriptions, Greek [<http://id.loc.gov/authorities/subjects/sh85066590>];

Inscriptions, Latin [<http://id.loc.gov/authorities/subjects/sh85066606>]

Coverage: Sicily (island) [<http://vocab.getty.edu/tgn/7030363>]; Sardegna, Isola di (island) [<http://vocab.getty.edu/tgn/7040284>]; Italian Peninsula (peninsula)

[<http://vocab.getty.edu/tgn/7023981>]; Classical world [<http://n2t.net/ark:/99152/p08m57hmxmp>]

Type: Dataset

17.

Title: Epigraphic Database Vernacular – EDV

Description: EDV is the first systematic collection of all the displayed documents in vernacular produced in Italy. It includes inscriptions dating from the 9th to the 15th cent., intended for any function – public or private – and performed on any surface (stone, plaster, canvas, fabric, glass, terracotta, metal, bone, etc.). All the inscriptions in a language other than Latin, or that show the intention, by the writer, to compose a text in vernacular are recorded.

Identifier: [[www.edvcorpus.com/wp](http://www.edvcorpus.com/wp)]

Creator: Università La Sapienza di Roma

Date: 2018

Format: text/html

Language: ita

Subject: Inscriptions [<http://id.loc.gov/authorities/subjects/sh85066566>]; Italian language–to 1300 [<http://id.loc.gov/authorities/subjects/sh85068807>]; Italian language–1300–1500 [<http://id.loc.gov/authorities/subjects/sh85068808>]

Coverage: Italy (nation) [<http://vocab.getty.edu/tgn/1000080>]; Middle Ages, 843–1517 [<http://n2t.net/ark:/99152/p06c6g3rhrz>]

Type: Dataset

18.

Title: Epigraphik-Datenbank Clauss / Slaby – EDCS

Description: The Epigraphik-Datenbank Clauss-Slaby (EDCS) is a searchable resource providing texts and bibliographic citations (lemmata of editions) for nearly all Latin inscriptions. It is edited by Manfred Clauss, and is the revised edition of a resource dating back to the late 1980. As of January 2018, EDCS contained texts for over 509,600 inscriptions previously published in print, together with over 112,000 images of inscriptions. Crosslinking to corresponding epigraphic records in 25 other databases (including EDR and EDH) is incorporated. The texts are simply regularized transcriptions drawn from previously published (print) editions.

Identifier: [<http://db.edcs.eu/>]

Creator: Universität Zürich; Katholische Universität Eichstätt-Ingolstadt

Date: 2018

Format: text/html

Language: deu; eng; spa; fra; ita

Subject: Inscriptions, Latin [<http://id.loc.gov/authorities/subjects/sh85066606>]

Coverage: Roman Empire (former nation/state/emire) [<http://vocab.getty.edu/tgn/7030347>]

Type: Dataset

19.

Title: Epigraphische Datenbank – EPIDAT

Description: The Database of Jewish epigraphy provides the inventory, documentation, editions and presentation of epigraphical collections. The geographical focus is on Germany, but inscriptions from Jewish cemeteries in The Netherlands, and also from Lithuania and the Czech Republic are also collected. The time span ranges from the 11th to the 20th century, from the Medieval and Early Modern periods to the Modern Era.

Identifier: [<http://www.steinheim-institut.de/cgi-bin/epidat>]

Creator: Salomon Ludwig Steinheim-Institute for German-Jewish History

Date: 2018

Format: text/html

Language: deu; eng

Subject: Jewish inscriptions [<http://id.loc.gov/authorities/subjects/sh85066602>]

Coverage: Germany (nation) [<http://vocab.getty.edu/tgn/7000084>]; Netherlands (nation) [<http://vocab.getty.edu/tgn/7016845>]; Lithuania (nation) [<http://vocab.getty.edu/tgn/7006542>]; Czech Republic (nation) [<http://vocab.getty.edu/tgn/1001780>]; Hochmittelalter [<http://n2t.net/ark:/99152/p0qhb66h8m4>]; Spätmittelalter [<http://n2t.net/ark:/99152/p0qhb66388g>]; Neuzeit [<http://n2t.net/ark:/99152/p0qhb669pgp>]

Type: Dataset

20.

Title: Europeana network of Ancient Greek and Latin Epigraphy – EAGLE

Description: The Europeana network of Ancient Greek and Latin Epigraphy is a best-practice network co-funded by the European Commission. EAGLE provides a single portal to the inscriptions of the Ancient World, by collecting, in a single readily-searchable database, more than 1.5 million items, currently scattered across 25 EU countries, as well as the east and south Mediterranean. The project makes available the vast majority of the surviving inscriptions of the Greco-Roman world, complete with the essential information about them and a translation into English.

Identifier: [<https://www.eagle-network.eu/>]

Creator: The EAGLE best practice network

Date: 2018

Format: text/html

Language: eng

Subject: Inscriptions, Greek [<http://id.loc.gov/authorities/subjects/sh85066590>];

Inscriptions, Latin [<http://id.loc.gov/authorities/subjects/sh85066606>]

Coverage: Classical World [<http://n2t.net/ark:/99152/p08m57hmxmp>]

Type: Dataset

21.

Title: The Glaser collection

Description: The collection of the Austrian scholar and explorer Eduard Glaser (1855–1908) was acquired in 1910 by the Academy of Sciences in Vienna. The epigrapher and specialist in the South-Arabian language brought back a huge amount of medieval Arabic manuscripts, and stone inscriptions, nowadays spread over Europe, as well as squeezes of the non-transportable ones, photographs, glass-negatives, diaries, and notes of historical importance. The Academy owns the latter precious documents of the 1880s and 1890s and within this project they are going to be digitally preserved and partly scientifically analysed.

Identifier: [<http://glaser.acdh.oeaw.ac.at/>]

Creator: Österreichische Akademie der Wissenschaften

Date: 2018

Format: text/html

Language: eng

Subject: Epigraphic South Arabian language [<http://id.loc.gov/authorities/subjects/sh98000742>]; Three-dimensional modelling [<http://id.loc.gov/authorities/subjects/sh2013001942>]

Coverage: Arabian Peninsula (general region) [<http://vocab.getty.edu/tgn/1012700>]; Arabian (culture) [<http://vocab.getty.edu/page/aat/300019797>]

Type: Dataset

22.

Title: Hesperia. Banco de datos de lenguas paleohispánicas

Description: The objective of the HESPERIA Paleohispanic Language Data Bank is the collection, organization and treatment of all the ancient linguistic materials related to the Iberian Peninsula (and those related to it from the South of France), with the exclusion of Latin, Greek and Phoenician inscriptions.

Identifier: [<http://hesperia.ucm.es/>]

Creator: Universidad Complutense de Madrid

Date: 2005

Format: text/html

Language: spa, eng

Subject: Inscriptions [<http://id.loc.gov/authorities/subjects/sh85066566>]; Coins [<http://id.loc.gov/authorities/subjects/sh85027797>]; Iberian language [<http://id.loc.gov/authorities/subjects/sh85063908>]; Celtiberian language [<http://id.loc.gov/authorities/subjects/sh96009143>]

Coverage: Iberian Peninsula [<http://vocab.getty.edu/tgn/7016676>]; Graeco-Iberian [<http://n2t.net/ark:/99152/p08m57h96rf>]; Edad del Hierro [<http://n2t.net/ark:/99152/p0qhb6666wx>]; Romano [<http://n2t.net/ark:/99152/p0qhb66r7np>]

Type: Dataset



23.

Title: I.Sicily

Description: I.Sicily is a project to create and make freely available online the complete corpus of inscriptions from ancient Sicily. The project includes texts in all languages (Greek, Latin, Phoenician/Punic, Oscan, Hebrew, and Sikel), from the first inscribed texts of the Archaic period (7th–6th centuries BC) through to those of late Antiquity (5th century AD and later). In the first instance the project is restricted to texts engraved on stone, but it is intended to expand that coverage in the future.

Identifier: [<http://sicily.classics.ox.ac.uk/>]

Creator: University of Oxford

Date: 2018

Format: text/html

Language: eng

Subject: Inscriptions, Greek [<http://id.loc.gov/authorities/subjects/sh85066590>]; Inscriptions, Latin [<http://id.loc.gov/authorities/subjects/sh85066606>]; Inscriptions, Phoenician [<http://id.loc.gov/authorities/subjects/sh85066622>]; Inscriptions, Punic [<http://id.loc.gov/authorities/subjects/sh85066626>]; Inscriptions, Oscan [<http://id.loc.gov/authorities/subjects/sh2001007060>]; Inscriptions, Hebrew [<http://id.loc.gov/authorities/subjects/sh85066592>]

Coverage: Sicily (island) [<http://vocab.getty.edu/tgn/7030363>]; arcaico [<http://n2t.net/ark:/99152/p0qhb665rrp>]; classic [<http://n2t.net/ark:/99152/p0qhb667rqt>]; romano [<http://n2t.net/ark:/99152/p0qhb66fq3k>]

Type: Dataset

24.

Title: Inscriptions of Aphrodisias – Iaph

Description: The aim of this online corpus is to present all the inscriptions found, on the site of Aphrodisias in Caria, or in its civic territory, up to the end of 1994. That provides a remarkable record of civic and personal life from at least the second century B.C. to at least the seventh century A.D., for the site is notably rich in inscriptions.

Identifier: [<http://insaph.kcl.ac.uk/iaph2007/index.html>]

Creator: King's College London

Date: 2007

Format: text/html

Language: eng

Subject: Inscriptions, Greek [<http://id.loc.gov/authorities/subjects/sh85066590>]; Inscriptions, Latin [<http://id.loc.gov/authorities/subjects/sh85066606>]

Coverage: Aphrodisias (deserted settlement) [<http://vocab.getty.edu/tgn/7002357>]; Roman-Early Byzantine Middle East (140 BC–AD 850) [<http://n2t.net/ark:/99152/p03wskdpzn9>]

Type: Dataset

25.

Title: Inscriptions of Israel/Palestine – IIP

Description: The Inscriptions of Israel/Palestine project collects and makes accessible all of the previously published inscriptions (and their English translations) of Israel/Palestine from the Persian period through the Islamic conquest (ca. 500 BCE–640 CE). There are about 15,000 of these inscriptions, written primarily in Hebrew, Aramaic, Greek and Latin, by Jews, Christians, Greeks, and Romans. They range from imperial declarations on monumental architecture to notices of donations in synagogues to humble names scratched on ossuaries.

Identifier: [<http://cds.library.brown.edu/projects/Inscriptions/>]

Creator: Brown University

Date: 2016

Format: text/html

Language: eng

Subject: Inscriptions, Hebrew–Palestine [<http://id.loc.gov/authorities/subjects/sh2009127267/>]; Inscriptions, Aramaic [<http://id.loc.gov/authorities/subjects/sh85066575/>]; Inscriptions, Greek [<http://id.loc.gov/authorities/subjects/sh85066590/>]; Inscriptions, Latin [<http://id.loc.gov/authorities/subjects/sh85066606/>]

Coverage: Israel (nation) [<http://vocab.getty.edu/tgn/1000119/>]; State of Palestine (autonomous area) [<http://vocab.getty.edu/tgn/7018359/>]; Jordan (nation) [<http://vocab.getty.edu/tgn/1000121/>]; Persian [<http://n2t.net/ark:/99152/p0qwcp63xkk/>]; Hellenistic [<http://n2t.net/ark:/99152/p0qwcp6wfdq/>]; Roman [<http://n2t.net/ark:/99152/p0qwcp6c8mc/>]; Byzantine [<http://n2t.net/ark:/99152/p0qwcp6m5m9/>]

Type: Dataset

26.

Title: Iscrizioni della Cirenaica greca - IGCyr; Iscrizioni metriche greche della Cirenaica – GVCyr

Description: The IGCyr corpus collects 920 inscriptions from the Greek Cyrenaica (VII–I century B.C.). The majority of these inscriptions have already been published, whereas 125 are unpublished. The GVCyr corpus contains 56 Greek metric inscriptions from Greek and Roman Cyrenaica, including 8 unpublished works.

Identifier: [<https://igcyr.unibo.it/>]

Creator: Alma Mater Studiorum Università di Bologna

Date: 2017

Format: text/html

Language: ita; eng; fra; ara

Subject: Inscriptions, Greek [<http://id.loc.gov/authorities/subjects/sh85066590/>]Coverage: Cyrenaica (historical region) [<http://vocab.getty.edu/tgn/7000643/>]

Type: Dataset

27.

Title: Linear B Electronic Resources – LiBER

Description: LiBER aims at producing an integrated database of Linear B documents, with the ultimate goal of providing scholars, and all those who are interested in the Mycenaean world, with an updated edition of the Linear B documents, along with a new set of search tools. Individual texts are supplied with transcriptions, critical apparatus, photographs as well as, whenever possible, with all the relevant information about findspots, scribes, chronologies, inventory numbers and places of preservation. The database can be searched by series of documents, syllabic sequences, logograms, scribes and findspots, while search results can be displayed both as lists of texts and interactive maps.

Identifier: [<http://liber.isma.cnr.it>]

Creator: CNR-Istituto di Studi sul Mediterraneo Antico

Date: 2013

Format: text/html

Language: eng

Subject: Inscriptions, Linear B [<http://id.loc.gov/authorities/subjects/sh85066610>]

Coverage: Greece [<http://vocab.getty.edu/tgn/1000074>]; Mycenaean [<http://n2t.net/ark:/99152/p08m57h97b5>]

Type: Dataset

28.

Title: The Neo-Babylonian Cuneiform Corpus – Nabucco

Description: NaBuCCo is a text-oriented website that aims at putting online textual metadata of an estimated 20,000 published Babylonian documentary sources including legal, administrative and epistolary records. The website collects all meta-textual data from the sources, paraphrases their content, makes the data available online, and links them (via partner websites) to the original source documents from which they are extracted. In addition to the text catalogue, the project offers a comprehensive up-to-date bibliography on Babylonia in the first millennium BCE.

Identifier: [<http://nabucco.arts.kuleuven.be>]

Creator: KU Leuven

Date: 2005

Format: text/html

Language: eng

Subject: Assyro-Babylonian [<http://id.loc.gov/authorities/subjects/sh85008838>];

Akkadian language—Texts [<http://id.loc.gov/authorities/subjects/sh2007100963>]

Coverage: Mesopotamia (general region) [<http://vocab.getty.edu/tgn/7001554>];

Neo-Assyrian/Babylonian Middle East (720–540 BC) [<http://n2t.net/ark:/99152/p03wskdxgnx>]

Type: Dataset

29.

Title: Online Corpus of Inscriptions of Ancient North Arabia – OCIANA

Description: A digital corpus of all known pre-Islamic inscriptions in North and Central Arabia provides a reading of each text both in roman transliteration and in fonts reproducing the ancient letters, together with a translation in English, references to earlier readings, commentary where necessary, bibliography, and all known information about the inscription (provenance, carving technique, relationship to other texts or to rock drawings, structures, etc.)

Identifier: [<http://163.1.184.24/fmi/webd/OCIANA>]

Creator: The Khalili Research Centre, University of Oxford

Date: 2014

Format: text/html

Language: eng

Subject: Inscriptions, Lihyanic [<http://id.loc.gov/authorities/subjects/sh85066608>];

Inscriptions, Safaitic [<http://id.loc.gov/authorities/subjects/sh85066631>];

Inscriptions, Thamudic [<http://id.loc.gov/authorities/subjects/sh85066637>]

Coverage: Arabian Peninsula (general region) [<http://vocab.getty.edu/tgn/1012700>];

Arabian (culture) [<http://vocab.getty.edu/page/aat/300019797>]

Type: Dataset

30.

Title: Open Richly Annotated Cuneiform Corpus – ORACC

Description: The Open Richly Annotated Cuneiform Corpus is an international cooperative project which provides facilities and support for the creation of free online editions of cuneiform texts and educational ‘portal’ websites about ancient cuneiform culture.

Identifier: [<http://oracc.museum.upenn.edu>]

Creator: University of Pennsylvania Museum of Anthropology and Archaeology

Date: 2017

Format: text/html

Language: eng

Subject: Cuneiform writing [<http://id.loc.gov/authorities/subjects/sh85034811>]

Coverage: Mesopotamia (general region) [<http://vocab.getty.edu/tgn/7001554>];

Bronze Age [<http://n2t.net/ark:/99152/p047fhmwtjz>]; Iron Age [<http://n2t.net/ark:/99152/p047fhmwx27>]

Type: Dataset

31.

Title: Packard Humanities Institute – PHI

Description: The Searchable Greek Inscriptions database (SGI) from Packard Humanities Institute contains Greek inscriptions from Greece including Crete, Cyprus, Thrace, the north coast of the Black Sea, Syria, Egypt, North Africa, Germany, and

unknown provenances organized by period and corpora. Inscriptions can be browsed by geographic area or searched for words and phrases.

Identifier: [<http://epigraphy.packhum.org>]

Creator: Packard Humanities Institute

Date: 2017

Format: text/html

Language: eng

Subject: Inscriptions, Greek [<http://id.loc.gov/authorities/subjects/sh85066590>]

Coverage: Greece (former nation/state/empire) [<http://vocab.getty.edu/tgn/7594735>]

Type: Dataset

32.

Title: Pondera

Description: The Pondera Online Project aims to collect and study ancient and medieval weights that were produced between the mid-sixth century BCE and the mid-fifteen century CE. Nowadays, more than 20,000 weights dating from these two millennia are registered, half of them from public and private collections, half of them from archaeological excavations.

Identifier: [<https://pondera.incal.ucl.ac.be/>]

Creator: UC Louvain

Date: 2018

Format: text/html

Language: eng

Subject: Weights and measures, Ancient [<http://id.loc.gov/authorities/subjects/sh85145970>]; Weights and measures, Medieval [<http://id.loc.gov/authorities/subjects/sh85145974>]

Type: Dataset

33.

Title: Projet Karnak

Description: The Karnak project aims to organize and to make accessible the textual documentation from the temples of Karnak. It is based on an exhaustive counting of documents and inscriptions collated on the original. Each document receives a unique identifier number when integrated into the database. All information relating to a document (typographic edition, transliteration, photographs, facsimiles, archival documents) can be accessed from a single notice.

Identifier: [<http://sith.huma-num.fr/karnak>]

Creator: LabEx Archimède - CNRS

Date: 2013

Format: text/html

Language: fra

Subject: Inscriptions, Egyptian [<http://id.loc.gov/authorities/subjects/sh85041341>]; Egyptian language [<http://id.loc.gov/authorities/subjects/sh85041339>]; Egyptian language–Writing, Hieroglyphic [<http://id.loc.gov/authorities/subjects/sh85041349>]  
 Coverage: Egypt (former nation/state/empire) [<http://vocab.getty.edu/tgn/7014986>]; Karnak (deserted settlement) [<http://vocab.getty.edu/page/tgn/7764757>]; First Intermediate Period Egypt (2168–2010 BCE/BC) [<http://n2t.net/ark:/99152/p03wskdxjnj>]; Middle Kingdom Egypt (2010–1640 BCE/BC) [<http://n2t.net/ark:/99152/p03wskdggq4n>]; Second Intermediate Period Egypt (1640–1548) [<http://n2t.net/ark:/99152/p03wskdzd99>]; New Kingdom Egypt (1548–1086) [<http://n2t.net/ark:/99152/p03wskddb3j>]; Third Intermediate Period Egypt (1086–664) [<http://n2t.net/ark:/99152/p03wskdmzfr>]; Late Period Egypt (664–332) [<http://n2t.net/ark:/99152/p03wskd47fw>]; Macedonian Egypt (332–304 BCE/BC) [<http://n2t.net/ark:/99152/p03wskdxnwr>]; Ptolemaic-Roman Egypt (304 BC–AD 640) [<http://n2t.net/ark:/99152/p03wskdftkm>]

Type: Dataset

34.

Title: Roman Inscriptions of Britain – RIB

Description: The website hosts volume one of The Roman Inscriptions of Britain, R.G. Collingwood's and R.P. Wright's magisterial edition of 2,401 monumental inscriptions from Britain found prior to 1955. It also incorporates all Addenda and Corrigenda published in the 1995 reprint of RIB (edited by R.S.O. Tomlin) and the annual survey of inscriptions published in Britannia since.

Identifier: [<https://romaninscriptionsofbritain.org/>]

Creator: Scott Vanderbilt

Date: 2018

Format: text/html

Language: eng

Subject: Inscriptions, Latin [<http://id.loc.gov/authorities/subjects/sh85066606>]

Coverage: Great Britain (island) [<http://vocab.getty.edu/tgn/7008653>]; Roman [<http://n2t.net/ark:/99152/p0gjgrs69ws>]

Type: Dataset

35.

Title: Runenproject Kiel

Description: The Kiel Rune Project was a scientific research project funded by the German Research Foundation (Deutsche Forschungsgemeinschaft - DFG) from 1993–1999 and from 2001–2012. Result of the project is a linguistic database of the oldest written attestations of the Germanic languages, the inscriptions in the Older Futhark. This database is meant to supplement the existing bibliographies, lexica and handbooks and to make the results of research in the field of runology available to

researchers working on the early stages of the Germanic languages and on Common Germanic.

Identifier: [[www.runenprojekt.uni-kiel.de](http://www.runenprojekt.uni-kiel.de)]

Creator: Kiel University

Date: 2012

Format: text/html

Language: deu, eng

Subject: Inscriptions [<http://id.loc.gov/authorities/subjects/sh85066566>]; Runes [<http://id.loc.gov/authorities/subjects/sh85115851>]

Coverage: Northern Europe [<http://vocab.getty.edu/tgn/4003757>]; Central Europe [<http://vocab.getty.edu/tgn/4003755>]; Germanic [<http://n2t.net/ark:/99152/p08m57hgq6p>]

Type: Dataset

36.

Title: Runische Schriftlichkeit in den germanischen Sprachen – RuneS

Description: The research project Runic Writing in the Germanic Languages (RuneS) investigates the oldest independently developed writing system in the Germanic languages, the runic script. The aim is to develop a system that will allow for the description of the inscriptions as text types. This means that the complete runic monument – the inscription-bearing object itself, the text written on it, accompanying iconographic elements and ornaments, the order of all these signs on the sign-bearing object as well as the historical circumstances of the find itself – need to be viewed in a synopsis, providing a basis for determining the function of each individual written document in the society it was produced in.

Identifier: [<http://www.runesdb.de>]

Creator: Akademie der Wissenschaften zu Göttingen

Date: 2018

Format: text/html

Language: deu, eng

Subject: Inscriptions [<http://id.loc.gov/authorities/subjects/sh85066566>]; Runes [<http://id.loc.gov/authorities/subjects/sh85115851>]

Coverage: Northern Europe [<http://vocab.getty.edu/tgn/4003757>], Central Europe [<http://vocab.getty.edu/tgn/4003755>]; Late Imperial [<http://n2t.net/ark:/99152/p06v8w4t5td>]; Middle Ages [<http://n2t.net/ark:/99152/p0pf7xr2szm>]

Type: Dataset

37.

Title: Scandinavian Runic-text Database

Description: The Scandinavian Runic-text Database aims to collect all Scandinavian runic inscriptions digitally. The inscriptions are published in transliterated and normalized form and with English translation.

Identifier: [www.nordiska.uu.se/forskn/samnord.htm]

Creator: Uppsala Universitet

Date: 1993

Format: text

Language: swe, eng

Subject: Inscriptions [http://id.loc.gov/authorities/subjects/sh85066566]; Runes [http://id.loc.gov/authorities/subjects/sh85115851]

Coverage: Northern Europe [http://vocab.getty.edu/tgn/4003757]

Type: Dataset

38.

Title: Sources of Early Akkadian Literature – SEAL

Description: SEAL aims to compile a complete indexed corpus of Akkadian literary texts from the 3rd and 2nd millennia BCE, relied on new collations and photos, attempting to enable the efficient study of the entire early Akkadian literature in all its philological, literary, and historical aspects.

Identifier: [http://www.seal.uni-leipzig.de]

Creator: Universität Leipzig; The Hebrew University of Jerusalem

Date: 2017

Format: text/html

Language: eng

Subject: Assyro-Babylonian [http://id.loc.gov/authorities/subjects/sh85008838]; Akkadian language—Texts [http://id.loc.gov/authorities/subjects/sh2007100963]

Coverage: Mesopotamia (general region) [http://vocab.getty.edu/tgn/7001554]; Akkadian Empire [http://n2t.net/ark:/99152/p0njrb4hqj5]; Old Babylonian/Assyrian Mesopotamia (2000–1600 BC) [http://n2t.net/ark:/99152/p03wskdr7sw]; Middle Assyrian [http://n2t.net/ark:/99152/p08tf6pjbv45]

Type: Dataset

39.

Title: Textdatenbank und Wörterbuch des Klassischen Maya – TWKM

Description: The goals of the project Interdisciplinary Dictionary of Classic Mayan are to provide a digital corpus of the texts and to compile a corpus-based dictionary of Classic Mayan. This dictionary will provide a comprehensive vocabulary of Classic Mayan and its use in writing.

Identifier: [http://mayawoerterbuch.de/]

Creator: University of Bonn

Date: 2014

Format: text/html

Language: eng

Subject: Inscription, Mayan [http://id.loc.gov/authorities/subjects/sh97007687]; Mayan languages—Writing [http://id.loc.gov/authorities/subjects/sh85082417]



Coverage: Central America (general region) [<http://vocab.getty.edu/tgn/7016739>]  
 Type: Dataset

40.

Title: Thesaurus Inscriptionum Raeticarum – TIR

Description: Thesaurus Inscriptionum Raeticarum (TIR) is an online edition of the Raetic inscriptions in the form of an interactive online platform of the MediaWiki type. The aim of the TIR project is a comprehensive collection, display and linguistic analysis of the inscriptions which are considered part of the Raetic corpus.

Identifier: [<http://www.univie.ac.at/raetica>]

Creator: Universität Wien

Date: 2016

Format: text/html

Language: eng

Subject: Inscriptions [<http://id.loc.gov/authorities/subjects/sh85066566>]; Raetian language [<http://id.loc.gov/authorities/subjects/sh87000364>]

Coverage: Alps (mountain system) [<http://vocab.getty.edu/tgn/7007746>]; Jüngere Eisenzeit [<http://n2t.net/ark:/99152/p0qhb6667kj>]

Type: Dataset

41.

Title: Trismegistos – TM

Description: Trismegistos is a conglomerate of databases, with Texts, Collections, Archives, People, Places, and Authors as main sections. It deals with texts from the ancient western world, dated between roughly 800 BC and AD 800. Its goal is to provide stable identifiers and general information about all texts for which there is physical evidence dated to that period.

Identifier: [<http://www.trismegistos.org>]

Creator: KU Leuven

Date: 2018

Format: text/html

Language: eng

Subject: Concordanances [<http://id.loc.gov/authorities/subjects/sh85030642>]

Type: Dataset

## Digital Lexica

1.

Title: KALAM - Word Analysis for Ancient South Arabian Languages

Description: KALAM reloaded is a linguistic text analyzing tool for the inscriptions written in closely related Semitic languages like Sabaic, Qatabānic, Minaic/Madhābian, and Ḥaḍramitic language. It is aimed at better resolving illegible passages/letters, while giving the full grammatical information with translation.

Identifier: [<http://kalam.ruzicka.net>]

Creator: Österreichische Akademie der Wissenschaften

Date: 2018

Format: text/html

Language: eng

Subject: Epigraphic South Arabian language [<http://id.loc.gov/authorities/subjects/sh98000742>]; Natural language processing (Computer science) [<http://id.loc.gov/authorities/subjects/sh88002425>]

Coverage: Arabian Peninsula (general region) [<http://vocab.getty.edu/tgn/1012700>]; Arabian (culture) [<http://vocab.getty.edu/page/aat/300019797>]

Type: Software

2.

Title: Lexicon of Greek Personal Names – LGPN

Description: The objective of LGPN is to collect and publish with documentation all known ancient Greek personal names (including non-Greek names recorded in Greek, and Greek names in Latin), drawn from all available sources (literature, inscriptions, graffiti, papyri, coins, vases and other artefacts), within the period from the earliest Greek written records down to, approximately, the sixth century A.D.

Identifier: [<http://www.lgpn.ox.ac.uk/>]

Creator: The British Academy; Oxford University

Date: 2014

Format: text/html

Language: eng

Subject: Names, Personal—Greek [<http://id.loc.gov/authorities/subjects/sh2010103095>]

Coverage: Greece (former nation/state/empire) [<http://vocab.getty.edu/tgn/7594735>]

Type: Dataset

3.

Title: The Pennsylvania Sumerian Dictionary – ePSD

Description: The PSD is preparing an exhaustive dictionary of the Sumerian language. It is designed as a corpus-based dictionary implemented with open-source software implementing XML-related standards.

Identifier: [<http://psd.museum.upenn.edu/>]

Creator: University of Pennsylvania Museum of Anthropology and Archaeology

Date: 2006

Format: text/html

Language: eng

Subject: Sumerian language [<http://id.loc.gov/authorities/subjects/sh85130413>]

Coverage: Mesopotamia (general region) [<http://vocab.getty.edu/tgn/7001554>];

Bronze Age [<http://n2t.net/ark:/99152/p047fhmwtjz>]; Late Babylonian [<http://n2t.net/ark:/99152/p08m57hp2rt>]

Type: Dataset

4.

Title: Ramses Online

Description: Ramses Online is a web interface giving access to some of the data and functionality of the annotated corpus of Neo-Egyptian texts developed at the University of Liège and known as the Ramses Project. It offers users a sub-corpus of Neo-Egyptian texts translated into French, all of whose occurrences are lemmatized and morphologically annotated.

Identifier: [<http://ramses.ulg.ac.be/>]

Creator: Université de Liège

Date: 2015

Format: text/html

Language: fra

Subject: Inscriptions, Egyptian [<http://id.loc.gov/authorities/subjects/sh85041341>];

Egyptian language [<http://id.loc.gov/authorities/subjects/sh85041339>]

Coverage: Egypt (former nation/state/empire) [<http://vocab.getty.edu/tgn/7014986>];

Second Intermediate Period Egypt (1640–1548) [<http://n2t.net/ark:/99152/p03wskdzd99>];

New Kingdom Egypt (1548–1086) [<http://n2t.net/ark:/99152/p03wskddb3j>];

Third Intermediate Period Egypt (1086–664) [<http://n2t.net/ark:/99152/p03wskdmzfr>]

Type: Dataset

5.

Title: Sabäisches Wörterbuch

Description: The DFG-funded project aims to create a Sabaic online dictionary. With about 6000 inscriptions dating from the 8th century BC to the 6th century AD, Sabaic is the best-tested dialect within the ancient South Arabian language community. In addition to extensive corpora of building, dedication and commemorative inscriptions, legal texts and a few hundred letters and economic texts, written on wooden sticks, are represented.

Identifier: [<http://sabaweb.uni-jena.de>]

Creator: Friedrich-Schiller-Universität Jena

Date: 2018

Format: text/html

Language: deu

Subject: Inscriptions, Sabaeen [<http://id.loc.gov/authorities/subjects/sh85066630>];

Sabaeen language [<http://id.loc.gov/authorities/subjects/sh98000731>]

Coverage: Arabian Peninsula (general region) [<http://vocab.getty.edu/tgn/1012700>];

Arabian (culture) [<http://vocab.getty.edu/page/aat/300019797>]

Type: Dataset

6.

Title: Thesaurus Linguae Aegyptiae – TLA

Description: A digital corpus of Egyptian (including Demotic) texts has been released to the public for computer-assisted search. Lemmatization and morpho-syntactic annotation of the text material allow for specific research from lexical, philological, linguistic, and historico-cultural points of view. All texts come with running translations to assist particularly non-specialists and scholars of neighbouring disciplines in their work. It is the purpose of the Thesaurus Linguae Aegyptiae to make available, in the form of a virtual dictionary, a tool for lexicographic research into the Egyptian language.

Identifier: [<http://aaew.bbaw.de/tla/>]

Creator: Berlin-Brandenburg Academy of Sciences and Humanities

Date: 2014

Format: text/html

Language: deu; eng

Subject: Inscriptions, Egyptian [<http://id.loc.gov/authorities/subjects/sh85041341>];

Egyptian language—Papyri [<http://id.loc.gov/authorities/subjects/sh85041342>];

Egyptian language [<http://id.loc.gov/authorities/subjects/sh85041339>]; Egyptian

language—Writing, Demotic [<http://id.loc.gov/authorities/subjects/sh85041347>]

Coverage: Egypt (former nation/state/empire) [<http://vocab.getty.edu/tgn/7014986>];

Old Kingdom Egypt (2670-2168 BCE/BC) [<http://n2t.net/ark:/99152/p03wskdgmtdf>];

First Intermediate Period Egypt (2168–2010 BCE/BC) [<http://n2t.net/ark:/99152/p03wskdxjnj>];

Middle Kingdom Egypt (2010–1640 BCE/BC) [<http://n2t.net/ark:/99152/p03wskdgg4n>];

Second Intermediate Period Egypt (1640–1548) [<http://n2t.net/ark:/99152/p03wskdgd99>];

New Kingdom Egypt (1548–1086) [<http://n2t.net/ark:/99152/p03wskddb3j>];

Third Intermediate Period Egypt (1086–664) [<http://n2t.net/ark:/99152/p03wskdmzfr>];

Late Period Egypt (664–332) [<http://n2t.net/ark:/99152/p03wskd47fw>];

Ptolemaic-Roman Egypt (304 BC–AD 640) [<http://n2t.net/ark:/99152/p03wskdftkm>]

Type: Dataset

7.

Title: Vocabulaire de l'Égyptien Ancien – VÉgA

Description: The Vocabulary of the Ancient Egyptian is an online digital dictionary. It groups and cross-check words, their attestations, their references, their various graphs in hieroglyphs, as well as the photographs of the texts concerned. This online tool is constantly updated by adding new words from unpublished and up-to-date sources based on the latest lexicographic studies.

Identifier: [<http://vega-vocabulaire-egyptien-ancien.fr/>]

Creator: LabEx Archimède - CNRS

Date: 2017

Format: text/html

Language: fra

Subject: Egyptian language [<http://id.loc.gov/authorities/subjects/sh85041339>]

Coverage: Egypt (former nation/state/empire) [<http://vocab.getty.edu/tgn/7014986>]

Type: Dataset

## Standards and Guidelines

1.

Title: EpiDoc guidelines

Description: EpiDoc is an international, collaborative effort that provides guidelines and tools for encoding scholarly and educational editions of ancient documents. In addition, the EpiDoc Website provides access to other tools and collaboration environments supported by the collaborative initiative.

Identifier: [<http://www.stoa.org/epidoc/gl/latest/>]

Creator: Tom Elliott; Gabriel Bodard; Elli Mylonas, Simona Stoyanova; Charlotte Tupman; Scott Vanderbilt

Date: 2007–2017

Format: text/html

Language: eng; ita; spa; bul

Subject: Digitization [<http://id.loc.gov/authorities/subjects/sh2002011497>];

Inscriptions [<http://id.loc.gov/authorities/subjects/sh85066566>]

Type: Text

2.

Title: Text Encoding Initiative – TEI

Description: The Text Encoding Initiative (TEI) is a consortium which collectively develops and maintains a standard for the representation of texts in digital form. Its chief deliverable is a set of Guidelines which specify encoding methods for machine-readable texts, chiefly in the humanities, social sciences and linguistics. Since 1994,

the TEI Guidelines have been widely used by libraries, museums, publishers, and individual scholars to present texts for online research, teaching, and preservation.

Identifier: [<http://www.tei-c.org>]

Creator: TEI Consortium

Date: 2007

Format: text/html

Language: eng

Subject: Digitization [<http://id.loc.gov/authorities/subjects/sh2002011497>]; Texts [<http://id.loc.gov/authorities/subjects/sh99001271>]

Type: Text

## Terminological Sources and Gazetteers

1.

Title: The Art & Architecture Thesaurus – AAT

Description: The AAT is a structured vocabulary containing terms and other information about concepts. Terms in AAT may be used to describe art, architecture, decorative arts, material culture, and archival materials. The target audience includes museums, libraries, visual resource collections, archives, conservation projects, cataloging projects, and bibliographic projects.

Identifier: [<http://www.getty.edu/research/tools/vocabularies/aat/>]

Creator: J. Paul Getty Trust

Date: 2017

Format: text/html

Language: eng

Subject: Subject headings [<http://id.loc.gov/authorities/subjects/sh85129426>]; Art–History [<http://id.loc.gov/authorities/subjects/sh85007488>]; Architecture [<http://id.loc.gov/authorities/subjects/sh85006611>]

Type: Dataset

2.

Title: Getty Thesaurus of Geographic Names Online – TGN

Description: TGN is a structured vocabulary containing names and other information about places. It is a thesaurus, compliant with ISO and NISO standards for thesaurus construction; it contains hierarchical, equivalence, and associative relationships. It is not a GIS (Geographic Information System): while many records in TGN include coordinates, these coordinates are approximate and are intended for reference only. The temporal coverage of the TGN ranges from prehistory to the present and the scope is global.

Identifier: [<http://www.getty.edu/research/tools/vocabularies/tgn/index.html>]

Creator: J. Paul Getty Trust

Date: 2017

Format: text/html

Language: eng

Subject: Gazetteers [<http://id.loc.gov/authorities/subjects/sh85053596>]

Type: Dataset

3.

Title: Graph of Dated Objects and Texts – GODOT

Description: The aim of this graph database system is to create and maintain a gazetteer of calendar dates in different calendar systems used in the Greek and Roman antiquity all across the Mediterranean sea. Like geographical gazetteers this authority list can be used to provide stable, unique identifiers (URIs) for each date in any of the calendar systems that has been used to refer to an astronomical day in any ancient source, be it papyri, ostraca or inscriptions. It will serve as a means to search and browse ancient texts by their precise temporal footprint using these URIs in digital editions and database or TEI/EpiDoc XML driven projects.

Identifier: [<https://godot.date>]

Creator: KU Leuven; King's College London; Heidelberg University; Heidelberg Academy of Sciences and Humanities

Format: text/html

Language: eng

Subject: Calendars [<http://id.loc.gov/authorities/subjects/sh85018851>]

Coverage: Classical World [<http://n2t.net/ark:/99152/p08m57hmxml>]

Type: Dataset

4.

Title: Iconclass

Description: Iconclass is a classification system designed for art and iconography. It is intended for description and retrieval of subjects represented in images (works of art, book illustrations, reproductions, photographs, etc.).

Identifier: [<http://www.iconclass.nl/home>]

Creator: Rijksbureau voor Kunsthistorische Documentatie

Date: 2012

Format: text/html

Language: eng

Subject: Classification [<http://id.loc.gov/authorities/subjects/sh85026719>]; Pictures [<http://id.loc.gov/authorities/subjects/sh85102012>]

Type: Dataset

5.

Title: Pelagios Peripleo

Description: Peripleo is a search service maintained by Pelagios Commons, that allows you to find community-curated content related to specific places.

Identifier: [<http://peripleo.pelagios.org/>]

Creator: Austrian Institute of Technology; Exeter University; Open University; University of London School of Advanced Study; Alexander von Humboldt Institute for Internet and Society.

Date: 2018

Format: text/html

Language: eng

Subject: Linked Data [<http://id.loc.gov/authorities/subjects/sh2013002090>]; Gazetteers [<http://id.loc.gov/authorities/subjects/sh85053596>]

Type: Interactive Resource

6.

Title: PeriodO - A Gazetteer of Period Definitions for Linking and Visualizing Data

Description: PeriodO is a public domain gazetteer of scholarly definitions of historical, art-historical, and archaeological periods. It eases the task of linking among datasets that define periods differently. It also helps scholars and students see where period definitions overlap or diverge.

Identifier: [<http://perio.do>]

Creator: University of Texas at Austin; University of North Carolina at Chapel Hill

Date: 2018

Format: text/html

Language: eng

Subject: Gazetteers [<http://id.loc.gov/authorities/subjects/sh85053596>]; Chronology, Historical [<http://id.loc.gov/authorities/subjects/sh85025412>]

Type: Dataset

7.

Title: Pleiades

Description: Pleiades is a community-built gazetteer and graph of ancient places. It publishes authoritative information about ancient places and spaces, providing services for finding, displaying, and reusing that information under open license. It publishes not just for individual human users, but also for search engines and for the widening array of computational research and visualization tools that support humanities teaching and research.

Identifier: [<http://pleiades.stoa.org/>]

Creator: Stoa Consortium; University of North Carolina at Chapel Hill; New York University

Date: 2017



Format: text/html

Language: eng

Subject: Gazetteers [<http://id.loc.gov/authorities/subjects/sh85053596>]

Type: Dataset

8.

Title: Standards for Networking Ancient Prosopographies – SNAP

Description: The SNAP:DRGN Graph is an authority list of ancient people, in the form of an aggregated triple store containing information about persons (or groups, gods, monsters, and other “person-like” entities from ancient sources) from the core project partners, from external data providers and newly created data and relationships between people and records. The triple store is available to researchers via a Sparql-endpoint and a RESTful API. The triple store can be browsed via URL or searched from a Sparql text page.

Identifier: [<https://snapdrgn.net>]

Creator: King’s College London; Duke University; KU Leuven; University of Southampton; Oxford University

Date: 2014

Format: text/html

Language: eng

Subject: Linked Data [<http://id.loc.gov/authorities/subjects/sh2013002090>];  
Prosopography [<http://id.loc.gov/authorities/subjects/sh86000287>]; Names,  
Personal—Greek [<http://id.loc.gov/authorities/subjects/sh2010103095>]; Names,  
Latin [<http://id.loc.gov/authorities/subjects/sh86004207>]

Type: Interactive Resource

## Portals and Websites

1.

Title: Arabia Antica

Description: Arabia Antica is the portal of Pre-Islamic Arabian studies conducted by the University of Pisa, Dipartimento di Civiltà e Forme del Sapere. It provides updates about epigraphic and philological projects, archaeological investigations, surveys in museums, international collaborations and publications by the research group based in Pisa, in addition to the state of art in this research domain.

Identifier: [<http://arabiantica.humnet.unipi.it>]

Creator: University of Pisa

Date: 2017

Format: text/html

Language: eng

Subject: Inscriptions [<http://id.loc.gov/authorities/subjects/sh85066566>], Epigraphic South Arabian language [<http://id.loc.gov/authorities/subjects/sh98000742>]; Semitic philology [<http://id.loc.gov/authorities/subjects/sh85119969>]; Archaeology (excavation) [<http://id.loc.gov/authorities/subjects/sh85046105>]; Archaeological museums and collections [<http://id.loc.gov/authorities/subjects/sh85006502>]  
 Coverage: Arabian Peninsula [<http://vocab.getty.edu/tgn/1012700>]  
 Type: Text

2.

Title: Hethitologie-Portal Mainz – HPM

Description: The Hethitologie-Portal Mainz is an open-access digital infrastructure for Hittitology and related fields of research in Ancient Near Eastern Studies. HPM gives access to an array of interconnected research documents, including critical editions of Hittite cuneiform texts, catalogues, bibliographies, onomastic databases as well as media archives with digital photos, drawings, and 3D models.

Identifier: [<http://hethiter.net>]

Creator: Die Akademie der Wissenschaften und der Literatur Mainz

Date: 2018

Format: text/html

Language: deu; eng

Subject: Inscriptions, Hittite [<http://id.loc.gov/authorities/subjects/sh85066594>]; Hittite language [<http://id.loc.gov/authorities/subjects/sh85061276>]

Coverage: Hittite Empire (former nation/state/empire) [<http://vocab.getty.edu/tgn/6002562>]; Late Bronze Age [<http://n2t.net/ark:/99152/p0m63njwrjx>]

Type: Text

3.

Title: Maya Decipherment

Description: Maya Decipherment is a weblog devoted to ideas and developments in ancient Maya epigraphy and related fields. It focuses on the dissemination and serious discussion of ideas related to Maya hieroglyphs and iconography, encompassing archaeology, linguistics, and other pertinent fields. Blog entries are categorized within one of five categories: Articles, Notes, Archives, News, and Books.

Identifier: [<https://decipherment.wordpress.com>]

Creator: David Stuart - University of Texas at Austin

Date: 2018

Format: text/html

Language: eng

Subject: Inscription, Mayan [<http://id.loc.gov/authorities/subjects/sh97007687>]; Mayan languages—Writing [<http://id.loc.gov/authorities/subjects/sh85082417>]

Coverage: Central America (general region) [<http://vocab.getty.edu/tgn/7016739>]

Type: Text

4.

Title: Munich Open access Cuneiform Corpus Initiative – MOCCI

Description: MOCCI has as a key objective the promotion of the digital humanities and easily accessible open-access data in order to widely disseminate, facilitate, and promote the active use and understanding of official inscriptions and archival texts of the Middle East in Antiquity, with an initial focus on those of ancient Mesopotamia (written in the cuneiform script and in the Akkadian and Sumerian languages), in academia and beyond. MOCCI seeks to create new and innovative ways for users to access the important and varied contents of numerous geo-referenced and linguistically-annotated editions of ancient records, primarily from the first millennium BC.

Identifier: [<http://www.en.ag.geschichte.uni-muenchen.de/research/mocci/index.html>]

Creator: Ludwig-Maximilians Universität München

Date: 2018

Format: text/html

Language: deu; eng

Subject: Cuneiform inscriptions, Akkadian [<http://id.loc.gov/authorities/subjects/sh85034804>]; Cuneiform inscriptions, Sumerian [<http://id.loc.gov/authorities/subjects/sh85034807>]

Coverage: Mesopotamia (general region) [<http://vocab.getty.edu/tgn/7001554>]; Bronze Age [<http://n2t.net/ark:/99152/p047fhmwjtjz>]; Iron Age [<http://n2t.net/ark:/99152/p047fhmwx27>]

Type: Text

5.

Title: Pelagios Commons

Description: Pelagios Commons is a community & infrastructure for Linked Open Geodata in the Humanities with the aim of linking historical materials through their common reference to particular places. Pelagios Commons offers online forums and real world events which allow anyone with an interest in connecting the past together to get involved. Moreover it provides tools and services which help people creating links and make use of them.

Identifier: [<http://commons.pelagios.org>]

Creator: Austrian Institute of Technology; Exeter University; Open University; Alexander von Humboldt Institute for Internet and Society.

Date: 2018

Format: text/html

Language: eng

Subject: Linked Data [<http://id.loc.gov/authorities/subjects/sh2013002090>]; Gazetteers [<http://id.loc.gov/authorities/subjects/sh85053596>]

Type: Text

6.

Title: Progetto Sinleqiunnini

Identifier: [www.pankus.com]

Description: Sinleqiunnini is a dedicated software designed for on-line edition of epigraphical sources and for the management of databases mainly concerned with cuneiform texts. The project has developed into a document management system which is able to process a variety of materials (transliterated, as well as normalized texts, photos etc.) and to perform a wide range of automatic functions while operating on different languages and syllabic scripts.

Creator: Università degli Studi di Napoli “L’Orientale”; Università Ca’ Foscari Venezia; CNR-Istituto di Studi sul Mediterraneo Antico

Date: 2013

Format: text/html

Language: eng

Subject: Cuneiform inscriptions [<http://id.loc.gov/authorities/subjects/sh85034803>];Inscriptions, Linear B [<http://id.loc.gov/authorities/subjects/sh85066610>]Coverage: Greece [<http://vocab.getty.edu/tgn/1000074>]; Mycenaean [<http://n2t.net/ark:/99152/p08m57h97b5>];Tell Mardikh (deserted settlement) [<http://vocab.getty.edu/tgn/7002266>];Ancient Syria (general region) [<http://vocab.getty.edu/tgn/8711750>];Early Bronze Age III [<http://n2t.net/ark:/99152/p0m63njtn97>];Late Bronze Age [<http://n2t.net/ark:/99152/p0cfv7g83cj>]

Type: Text; Dataset

# Appendix B

## Mapping of Selected Concepts from the Index

The present section contains selected terms from the Index of Concepts, with mapping to the Library of Congress Subject Headings and the Getty Art and Architecture Thesaurus for their disambiguation and definition.

### *affixes*

Grammar, Comparative and general—Affixes [<http://id.loc.gov/authorities/subjects/sh85056264>]

### *annotation*

Annotations [<http://id.loc.gov/vocabulary/ethnographicTerms/afset000606>]

### *assimilation*

Assimilation (Phonetics) [<http://id.loc.gov/authorities/subjects/sh85008792>]

### *authority*

Authority files (Information retrieval) [<http://id.loc.gov/authorities/subjects/sh85009792>]

### *capture*

Electronic data processing—Data entry [<http://id.loc.gov/authorities/subjects/sh85042291>]

### *carrier*

Carriers (physical media) [<http://id.loc.gov/vocabulary/ethnographicTerms/afset002790>]

### *collation*

Collations [<http://vocab.getty.edu/aat/300311715>]

### *commentary*

Commentaries [<http://vocab.getty.edu/aat/300026098>]

### *concordance*

Concordances [<http://id.loc.gov/authorities/subjects/sh85030642>]

### *controlled terms*

Controlled vocabularies [<http://id.loc.gov/authorities/genreForms/gf2014026070>]

*conversion*

File conversion (computer science) [<http://id.loc.gov/authorities/subjects/sh92005723>]

*curation*

Data curation [<http://id.loc.gov/authorities/subjects/sh2015001855>]

*digital preservation*

Digital preservation [<http://id.loc.gov/authorities/subjects/sh95004496>]

*disambiguation*

Disambiguation [<http://id.loc.gov/vocabulary/ethnographicTerms/afset005248>]

*encoding*

Encoding [<http://id.loc.gov/vocabulary/ethnographicTerms/afset006068>]

*execution*

Execution (artistic concept) [<http://vocab.getty.edu/aat/300069715>]

*formulae*

Formulaic expressions [<http://id.loc.gov/vocabulary/ethnographicTerms/afset007344>]

*harvest*

Metadata harvesting [<http://id.loc.gov/authorities/subjects/sh2007001751>]

*host*

Web hosting [<http://id.loc.gov/authorities/subjects/sh2002010032>]

*infixes*

Infixes [<http://id.loc.gov/authorities/subjects/sh00006127>]

*interoperability*

Networking (Telecommunication) [<http://id.loc.gov/authorities/subjects/sh94007902>]

*lacunae*

Lacunae [<http://vocab.getty.edu/aat/300263354>]

*ligature*

Ligatures (script forms) [<http://vocab.getty.edu/aat/300195902>]

*Linked Data*

Linked Data [<http://id.loc.gov/authorities/subjects/sh2013002090>]

*maintenance*

Maintenance [<http://id.loc.gov/authorities/subjects/sh85079931>]

*manuscript*

Manuscripts [<http://id.loc.gov/authorities/subjects/sh85080672>]

*mapping*

Concept mapping [<http://id.loc.gov/authorities/subjects/sh2007007421>]

Metadata crosswalk [<http://id.loc.gov/authorities/subjects/sh2010013649>]

*metadata*

Metadata [<http://id.loc.gov/authorities/subjects/sh96000740>]

*migration*

System migration [<http://id.loc.gov/authorities/subjects/sh93000342>]

*mining*

Data mining [<http://id.loc.gov/authorities/subjects/sh97002073>]

*object-oriented language*

Object-oriented programming languages [<http://id.loc.gov/authorities/subjects/sh2006006405>]

*ontology*

Ontologies (Information retrieval) [<http://id.loc.gov/authorities/subjects/sh2005006014>]

*Open Access*

Open access publishing [<http://id.loc.gov/authorities/subjects/sh2005002533>]

*open source*

Open source software [<http://id.loc.gov/authorities/subjects/sh99003437>]

*platform*

Computing platforms [<http://id.loc.gov/authorities/subjects/sh2011003111>]

*portal*

Web portals [<http://id.loc.gov/authorities/subjects/sh99005116>]

*prefix*

Suffixes and prefixes [<http://id.loc.gov/authorities/subjects/sh2001009073>]

*query*

Querying (Computer science) [<http://id.loc.gov/authorities/subjects/sh2005008252>]

*repository*

Institutional repository [<http://id.loc.gov/authorities/subjects/sh2006003967>]

*retrieval*

Information retrieval [<http://id.loc.gov/authorities/subjects/sh85066148>]

*root*

Roots [<http://id.loc.gov/authorities/subjects/sh00007663>]

*semantic web*

Semantic Web [<http://id.loc.gov/authorities/subjects/sh2002000569>]

*semantics*

Semantics [<http://id.loc.gov/authorities/subjects/sh85119870>]



# List of Figures and Tables

- Figure 1.1:** CSAI homepage (2010) — 2
- Figure 1.2:** DASI data entry interface — 5
- Figure 1.3:** Early Sabaic boustrophedon inscription with monogram and symbols (MŞM 149) — 7
- Figure 1.4:** DASI XML editor — 8
- Figure 1.5:** DASI Lexicon — 14
- Figure 2.1:** Screenshot of the basic (= *find*) information on the Aspa stone (Sö 137), cf. [runesdb.eu/find-list/d/fa/q////6/f/4782/] (for an overview of the runic objects of our corpus see [runesdb.eu/find-list]) — 27
- Figure 2.2:** Random sample of snippets of the *þ*-rune with the assigned graph-type variants and graph types — 30
- Figure 2.3:** Highly simplified structure of the graphemic section of the database — 30
- Figure 2.4:** Simplified view of bilingual data in the database — 32
- Figure 2.5:** The graphemic data input mask — 34
- Figure 3.1:** Home page of Hesperia — 41
- Figure 3.2:** Home page of the AELAW Database — 46
- Figure 4.1:** Synoptic scheme of parallel hierarchies — 53
- Figure 5.1:** Map of the Yucatan peninsula with major archaeological sites (drawing by N. Grube and U. Lohoff-Erlenbach) — 66
- Figure 5.2:** Stela D from the Maya site of Pusilha, Belize, with references to local dynastic and political history (drawing by C. Prager) — 68
- Figure 5.3:** Examples of basic sign functions in Maya writing (concept by C. Prager) — 69
- Figure 5.4:** Graphotactic patterns in Maya writing. a) Syllabic spellings of *y=uk'ib* “his drinking vessel”. b) Different spellings of *K'inich*, the proper name of the Classic Maya sun god. c) Scribal plays for the word *kakaw* “cacao”. d) Two spellings of the word *u=pakbu tuunil* “his stone-lintel” — 70
- Figure 5.5:** Overview of the ontology-based metadata schema for describing artefacts and their contexts — 74
- Figure 5.6:** HTML and JavaScript-based entry mask to record the metadata in TextGrid — 76
- Figure 5.7:** Allographic spellings of the sign 595 for the syllable no. a) Full form of 595, represented by the graph 595tv. b-c) Sign 595 in the spelling ko-ko=no=ma < kok-n-om “guardian” d) 595 used in the word TZUTZ=no=ma < tzutz-n-om “planter”, e) CHOK=no=ma < chok-n-om “scatterer”, f) yu-ku=no=ma < yuk-n-om “shaker” (drawings by Stephen Houston, Linda Schele) — 77
- Figure 5.8:** Graphs of sign 528 representing the syllable ku, the morphograph TUN “stone”, and the day sign CHAHUK, the name of 19th day in the Maya calendar (drawings by M. Zender) — 78
- Figure 5.9:** Example evaluation of the transliteration value “TUN” for Sign No. 528 — 79
- Figure 7.1:** Home menu of the CIP data bank, with basic access links — 94
- Figure 7.2:** Top: Searching by countries layout. Below: List of inscriptions layout — 96
- Figure 7.3:** Inscription main layout, integrating data from various primary and secondary tables — 98
- Figure 7.4:** Top: Line of text file (Basic layout). Below: List of find-spots (example of layout integrating data from secondary tables) — 99

- Figure 7.5:** Graphic material main layout — 100
- Figure 8.1:** Dadanitic inscriptions at al-‘Udhayb (al‘Ulā, Saudi Arabia). (Photograph by C.J. Robin) — 103
- Figure 8.2:** Detail of Fig. 8.1, Dadanitic inscriptions at al-‘Udhayb (al‘Ulā, Saudi Arabia). (Inscriptions U 011–019, 021–026, see OCIANA). (Photograph by C.J. Robin) — 104
- Figure 8.3:** Safaitic inscriptions at Jabal Says, southern Syria (C 25–32, see OCIANA). (Photograph by M.C.A. Macdonald) — 105
- Figure 8.4:** Safaitic inscriptions on a stone at al-‘Īsāwī, southern Syria (C 3260–3264 see OCIANA). (Photograph by M.C.A. Macdonald) — 105
- Figure 9.1:** Result for lemma. The example is *sb’* — 127
- Figure 10.1:** Altar inscription, Wukro, Ethiopia (photo by R. Ruzicka) — 133
- Figure 10.2:** Squeeze sample collected by Eduard Glaser — 134
- Figure 10.3:** Analyzing *ybn* using KALAM — 137
- Figure 10.4:** Analyzing *stšr* using KALAM — 138
- Figure 10.5:** Using the dictionary. If one enters “*mḥfd*” selecting “Sabaic”, “Minaic” and “use dictionary” one obtains “tower”. Similarly, entering “tower” and selecting “English word” produces “*mḥfd*” — 138
- Figure 10.6:** Connecting KALAM and SabaWeb. We take again “*stšr*”, “use dictionary”, “Sabaic only” and select “SabaWeb”. Clicking on the SabaWeb link, all additional information is shown — 139
- Figure 10.7:** Result page from SabaWeb — 139
- Figure 11.1:** ARMEP base map showing the find-spots of ancient texts — 147
- Figure 11.2:** Example of ARMEP filter by metadata results — 148
- Figure 11.3:** Example of ARMEP filter by content results — 149
- Figure 11.4:** Sample “Cluster Overview” (left) and “Item View” (right) — 149
- Figure 12.1:** Main axis of the temple of Amun-Ra at Karnak (© CNRS-CFEETK) — 156
- Figure 12.2:** A scene and its inscriptions from the White Chapel of Senusret I (ca. 2000 BCE) — 158
- Figure 12.3:** Orthophotographic survey, data processing in Photoscan and detail of orthophotography of an inscription several meters high acquired by means of this technique — 160
- Figure 12.4:** The world *nsyt* « Kingship » in the inscriptions of Karnak — 162
- Figure 13.1:** The frontpage of HPM — 168
- Figure 13.2:** The WebGLViewer of HPM allows the collation of cuneiform tablets in the web browser and provides several tools for measuring and enhancement — 173
- Figure 13.3:** Hittite text with mark-up in LibreOffice — 176
- Figure 13.4:** An example for the reuse of an older manuscript as database by tagging it with styles — 176
- Figure 13.5:** On mouse-over the SVG join sketch displays the publication number of the fragment and a link to the relevant photos — 178
- Figure 14.1:** Number of inscriptions per region (in red) and the percentage the region occupies in the sample (in blue) — 183
- Figure 14.2:** Number of inscriptions per region (according to the vernacular used) — 184
- Figure 14.3:** Use of vernacular in Central and Southern Italy — 185
- Figure 16.1:** Diagram of PeriodO data model — 207

- Figure 16.2:** Using the PeriodO reconciler with OpenRefine to match period terms from the EDH search page to period definitions in the gazetteer — 211
- Figure 17.1:** Main terms in different languages in the Type of Inscription EAGLE Vocabulary — 220
- Figure 18.1:** Epitaphs and files (*per anno*) — 232
- Figure 18.2:** Development of upper part forms of headstones — 234
- Figure 18.3:** Buckled Shoe, Frankfurt 1795 — 238
- Figure 19.1:** Graphic representation of the data organisation of I.Sicily — 244
- Table 8.1:** Some of the fields assigned to each inscription in the database, with links to other tables within OCIANA — 111

## Index

### Index of Ancient and Modern Regions and States

- Abruzzi 184  
 Africa 196  
 Anatolia 167, 172, 179  
 Apulia 183-184  
 Arabia / Arabian Peninsula XIV, 1, 3, 6, 12, 15, 102-104, 136  
 Armenia 142  
 Assyria 141-144, 151  
 Babylonia 141-143, 145, 168  
 Basilicata 184  
 Belize 65, 68  
 Calabria 184  
 Campania 184  
 Central and East Europe 231  
 Central Italy 184  
 Czech Republic 232  
 Egypt 84, 87, 155, 161, 163, 193-196  
 Eritrea 84-88  
 Ethiopia 84-85, 87-88, 103, 133  
 France 38-39, 46, 157  
 Friuli 184  
 Gaul 37  
 Germany 167, 171, 231-232, 237  
 Great Britain 25  
 Guatemala 65  
 Hispania 37, 39, 46, 217  
 Honduras 65  
 Iberian Peninsula 38  
 Iran 142  
 Iraq 103  
 Italy XVI, 46, 180-185, 190, 205, 237  
 Jordan 103-104, 115  
 Lazio 181-183, 188  
 Lithuania 232  
 Lombardia 183  
 Mediterranean XV, 3, 39, 49, 93, 220  
 Mesopotamia 142-143, 146  
 Mexico 65  
 Middle East XVI, 104, 107, 141  
 Near East XIV, 13, 54  
 North Arabia XV, 102, 104, 106-108, 115  
 Northern Italy 46, 184  
 Oman 1, 103  
 Sardinia 184  
 Saudi Arabia 1, 103-104, 114-115  
 Scandinavia 23  
 Sicily XVII, 46, 183-184, 240, 246, 249-250  
 South Arabia / Southern Arabia 9, 11, 14, 85, 103, 129, 133  
 Southern Italy 184-185  
 Sudan 84  
 Syria 49, 55, 103-106, 115  
 The Netherlands 232  
 Turkey 142, 167, 206  
 Tuscany 182-184  
 Veneto 181-184  
 Yemen 1, 84-85, 102-104, 115, 120

## Index of Languages and Scripts

- Akkadian 53, 103, 142-146, 148, 151, 168  
 Amharic 103  
 Ancient Egyptian 155, 157, 161-164, 174  
 Ancient North Arabian XIV, 3, 9, 11, 13, 102-103, 115-116  
 Ancient South Arabian (*see also* Epigraphic South Arabian, South Arabian) XIV, XVI, 1-3, 9, 11, 13, 102-103, 115, 118-121, 129, 133, 135, 139  
 Anglo-Frisian fuþorc 22, 26  
 Aquitanian 38, 46  
 Arabic 110, 113, 120, 124, 126, 129, 133, 136, 150, 163  
 Aramaic XIV, 3, 9, 11, 13, 103, 150, 152, 195  
 Balcanic 46  
 Camunic 47  
 Catalan 182  
 Celtiberian 38-39, 42-43, 46  
 Celtic 38-39, 46, 195  
 Chamito-Semitic XV  
 Classic Mayan XV, 65, 67, 71-73, 80-81, 253  
 Classical Ethiopic (*see also* Ethiopic, Gəʿəz, Old Ethiopic) 81  
 cuneiform XVI, 49-50, 52-56, 59-60, 62, 141-144, 148, 152, 167-170, 172-174, 177-179, 253  
 Dadanitic 103-104, 106-107, 113  
 Demotic XVI, 155, 157-159, 161, 163, 194, 196-197  
 Early Italian (*see also* vernacular) 190  
 Eblaite 58-59, 168  
 Elamite 142-143, 146  
 Elymian 46, 241  
 English 21, 24, 26, 27, 31-32, 34, 62, 107, 110, 138, 142-143, 145-146, 151, 163, 219  
 Epigraphic South Arabian (*see also* Ancient South Arabian, South Arabian) 118  
 Ethiopic (*see also* Classical Ethiopic, Gəʿəz, Old Ethiopic) 84-86, 88, 90, 129  
 Etruscan 42, 47, 195  
 Faliscan 46  
 fidal (*see also* Ethiopic) 81, 84, 88-89  
 Frisian 27  
 Gallo-italic vernaculars 184  
 Gaulish 38, 42, 46, 195  
 Gəʿəz (*see also* Ethiopic) 85, 88, 90-91, 103  
 German 21, 24, 26, 31-32, 34, 126, 129-130, 142, 146, 219, 237-238  
 Germanic 15, 21  
 Greek XVII, 36-39, 46-47, 84-86, 88, 90-91, 103, 195-198, 202, 208, 213, 216-217, 225-226, 241, 247, 254  
 Greek Ionic 39  
 Ḥadramitic 2, 129, 133, 135-136, 139  
 Hasaitic 103  
 Hebrew 150, 231-232, 235, 237, 241  
 hieratic XVI, 155, 157-159, 161, 163  
 hieroglyphic (Egyptian) XVI, 155, 157, 159, 161, 163, 194  
 hieroglyphic (Luwian) 54, 174  
 hieroglyphic (Maya) 65, 67, 69, 71, 72, 76, 78-81  
 Himaitic 115  
 Hismaic 107, 115-116  
 Hittite 167, 170-174, 176, 178  
 Iberian 38-39, 42-44, 46, 195  
 Imperial Aramaic 103  
 Indo-European XV, 38, 40  
 Italic 195  
 Late Egyptian 158  
 Latin XVII, 36-40, 42, 46-47, 58, 103, 120, 180-181, 184, 190, 195, 197, 216, 219, 241, 254  
 Lepontic 38, 46  
 Ligurian 46  
 Linear B 49-50, 52-54, 58, 253  
 Lusitanian 38-40, 42, 46  
 Luwian 152  
 Messapian / Messapic 46, 195  
 Middle Egyptian 158  
 Minaic 2, 11, 118, 129, 133, 135-136, 138-139  
 Modern South Arabian 129  
 Nabataean 2, 103  
 North-West Semitic 103  
 Northwestern Iberian 43  
 Ogham 195  
 Old Aramaic 103  
 Old English 24-25, 35  
 Old Ethiopic (*see also* Classical Ethiopic, Ethiopic, Gəʿəz) 81  
 Old French 182  
 Old Persian 142-143  
 older fuþark 22, 26, 35  
 Oscan 42, 47, 241  
 Palaeo-European 45-47  
 Palaeohispanic XV, 36, 38-43, 45, 254

- Palmyrene 103, 209  
 Phoenician XV, 2, 36-38, 42, 46-47, 93-94,  
 100, 241  
 Phoenico-Aramaic 103  
 Ptolemaic 158  
 Punic XV, 42, 93-94, 195, 241  
 Qatabanic 2, 118, 129, 133, 135-136, 139  
 Raetic 47, 195  
 runic XV, 21-28, 31, 34-35, 195, 253  
 Sabaean / Sabaic XVI, 2, 7, 11, 85, 88, 102,  
 118-126, 129, 133, 135-139  
 Sabellic 46  
 Safaitic 102, 104-108, 110-111, 115-117, 276,  
 294  
 Semitic XIII, XIV, 4, 9, 14, 85, 88, 103, 113,  
 119-121, 129, 136, 195, 254  
 Sican 46  
 Sikel 46, 241  
 South Arabian (*see also* Ancient South Arabian,  
 Epigraphic South Arabian) 1-3, 84-86,  
 118, 199  
 South Germanic 27  
 South Semitic 103  
 Southern Iberian 42  
 Southwestern language (*see also* Tartessian)  
 38-39, 42, 46  
 Sumerian 142-146, 150-151, 168  
 Tartessian (*see also* Southwestern language)  
 38, 46  
 Taymanitic 107  
 Thracian 46  
 Umbrian 42  
 Urartian 142-143  
 Vasconic 38-39, 46  
 Venetic 46-47  
 vernacular (Italian) XVI, 180-182, 184-185  
 younger fup̄ark/fup̄ork 22, 26-27, 35

## Index of Concepts (Normalized)

- 3D 167-170, 172, 174-175, 203, 217
- abbreviation 6, 26, 77, 91, 120, 197, 221
- affixation 69
- affixes 9
- aggregation XVIII, 12, 203-204, 209, 212-213, 218-219, 226, 250
- aggregator XVII, 12, 216-217, 224-226, 236
- alignment (*see also* mapping, harmonization) 10, 219, 242-243
- aligned 184, 219, 224, 226, 243, 248-249
- allograph (ic notation) (*see also* variant) 25, 67, 72, 77, 79, 253
- alphabet 38-40, 42, 103, 115
- alphabetic scripts/writing systems XIV, XV, 52
- ambiguity (*see also* disambiguation, uncertain interpretation, uncertain reading) 14, 56, 254
- ambiguous forms XVI, 118, 124
- annotation (*see also* mark-up, tagging) XVI, 1, 4-5, 8, 49, 50, 54-56, 59-62, 73, 80-81, 88-90, 92, 118-119, 170, 177, 245, 247, 253
- annotated XV, XVII, XVIII, 3, 7, 11, 14, 49, 52, 76, 80, 88-89, 119, 133, 141-142, 162, 168, 198
- anthropology 181
- anthroponyms (*see also* names of individuals, personal name) 122, 161
- API 225, 244, 247
- apparatus criticus (*see also* commentary, critical apparatus) 5, 7-9, 107, 110-111, 113, 253
- archaeology XIII, 26, 159, 204, 223
- archaeological context (*see also* material context) 26, 42, 52, 73, 203, 213
- archaeological datasets XVII, 254
- archaeological object (*see also* artefact, carrier, support, text-bearing object) 169, 233
- archaeologists XIV, 44, 205
- architecture XVII, 11, 50-52, 54, 56-57, 59, 72, 74, 95, 147, 181, 190, 233
- art history (*see also* history of art) 26, 204, 231, 233
- art-historians XIV, 181, 231
- artefact (*see also* archaeological object, carrier, support, text-bearing object) 3, 11, 73-74, 85, 87, 190
- assimilation (*see also* non-assimilated forms) 136
- attestation XV, 12, 71, 116, 123, 126, 130, 161-163, 196, 199, 254
- authority 60, 206, 208, 211, 237, 242-243, 245-249, 251, 255
- automated/automatic annotation/mark-up/tagging 81, 140, 226
- autopsy / autoptic reading 26, 157, 242-243, 245-246, 249, 256
- best practice XIV, XVII, 3-4, 15, 255, 257
- bibliography (*see also* literature) XIII, 5, 13, 95, 97-98, 107-108, 115, 145, 167, 172, 187-188, 194, 242, 244-245
- bibliographic(al) references/citation 3, 40-41, 43, 47, 60, 73, 112, 172, 241
- bilingual terminology 24, 31-32, 34
- bilingual inscriptions 143
- calligraphy 67, 79
- capture (*see also* import, incorporation) 203, 243, 246
- carrier (*see also* archaeological object, artefact, support, text-bearing object) 12, 23, 254
- cartographic distribution (*see also* geographical distribution, spatial distribution) 44
- chronology (*see also* date, period, time) 5, 13, 86, 162, 202, 206, 213, 233
- CIDOC CRM 77, 216, 226, 238, 249, 253
- clitic 9
- codification (*see also* encoding) 42-43
- coin XV, 37, 39, 41, 44, 204, 209
- collation 95, 173
- commentary (*see also* apparatus criticus, critical apparatus) 35, 40, 52, 81, 107, 110-111, 113, 128, 187
- conceptual model/schema/scheme (*see also* data model, modelling) 1, 5, 56-57
- concordance (*see also* siglum) 90, 93, 95, 108, 110, 112
- consistency 23, 52, 57, 177, 223
- content provider 221
- contextual information 26, 58-59, 254
- controlled terms (*see also* thesauri, vocabulary) 4, 10
- conversion 1, 10, 57, 85, 89, 194, 219, 242-244
- converted 108, 121, 175-176, 198, 220-221, 243
- co-occurrence 50, 62, 198
- cooperative annotations 59
- coordinates (geographic) 13, 44, 151-152
- copyright 28, 174, 256

- corpus / corpora XIII-XV, 1-2, 6, 11, 13-14, 21, 26-27, 29, 40, 45, 49-50, 52-53, 74, 78-79, 84, 87-88, 93-95, 97, 100, 102, 106-112, 114-116, 118-119, 121, 126, 128, 140-148, 150, 152, 155-157, 161, 163, 168, 179, 193, 195-196, 198, 203, 208-209, 212-213, 242-243, 251, 254, 256
- correction 7, 106, 151, 177, 190, 197, 253
- Creative Commons 110, 163, 237
- critical apparatus (*see also* apparatus criticus, commentary) 37, 187, 245-246
- critical edition XIV, 46, 58, 93-94, 186-188
- cultural heritage XIV, 6, 217, 231
- curation 218, 236, 246
- data entry 4-5, 7, 32-33, 109, 223, 246
- data model (*see also* conceptual model, modelling) 1, 5-6, 12, 49-50, 58, 61-62, 77, 87, 207, 224, 226, 236, 238
- data persistency (*see also* digital preservation) 57
- database / databank XV, XVI, XVII, 1, 3-6, 21-38, 40-52, 57, 59, 61-63, 65, 71-73, 75-76, 81-82, 85, 93-95, 98, 100, 102, 106-112, 114-116, 119, 130, 133, 144, 159, 167-169, 171-172, 174-176, 178-180, 183, 185-187, 190, 193-199, 202-203, 207-208, 211-214, 217, 219, 221, 223-226, 229, 231-232, 236, 242-245, 247-248, 253-254
- date (*see also* chronology, time) 39, 73, 93, 97, 111, 114-115, 144, 147, 161, 168, 172, 186, 188, 190, 197, 199, 205-206, 208-209, 211, 220, 222, 235, 242, 250-251
- dating 11, 23, 25-26, 35, 177, 205, 208, 243
- date-range (*see also* period) 208
- decipherment 71-73, 78, 81, 107-108, 114, 134, 179
- deciphered XV, XVII, 36, 38-39, 43, 65, 69, 81, 86, 106-107, 115, 254
- defective orthography/script 120, 122-123
- determinative 53, 55-56
- diacritic(al) characters/marks/signs/symbols 4, 9, 42-43, 78, 110, 135
- dictionary XV, XVI, 14, 35, 57-58, 65, 71-72, 81, 93, 102, 106, 108, 110, 112, 115, 118-119, 121-126, 128-131, 133, 138-139, 150, 161, 163, 178-179, 253-254
- digital edition (*see also* electronic edition) 1, 7, 9, 37, 50, 52, 62, 226-227, 230-231, 236, 255
- digital heritage XIII, 16, 140
- digital humanities XIII, 3, 16-17, 61, 65, 82, 112, 141-142, 167, 180, 184, 214, 236, 238
- digital library 168, 210, 236
- digital preservation (*see also* data persistency) 256
- diplomatic (edition) 90-91, 181, 187-188
- direction of script (*see also* text direction) 9
- disambiguation XVII, 123-124, 150-151, 204
- dissemination XIV, XVIII, 2, 15, 49, 159, 162, 173
- divine name (*see also* god names, names of gods, theonym) 110, 113-114, 172
- drawing (*see also* graphic documentation, graphic material, visual documentation) 37, 43, 66, 68, 77-78, 98, 167-169, 174, 177
- Dublin Core XVII, 210
- EDM 12, 216, 236
- electronic edition (*see also* digital edition) 95
- encoding (*see also* codification, mark-up, tagging) XIV, XV, XVI, XVII, XVIII, 1, 3-12, 14-15, 19, 36, 47, 49-52, 54, 62, 79-80, 84, 87-88, 91-92, 109, 115, 121, 157, 161, 174, 187, 202, 212, 220-222, 226, 233, 236, 238, 253-256
- EpiDoc 4, 7-8, 10, 12, 84, 88, 91, 109, 162, 187, 202-203, 212-213, 216-217, 219-224, 226-227, 233, 237, 240, 242-251, 253, 255
- erasure 53
- ethnic (name) 161, 163, 194
- execution 29, 187, 242
- export 88, 217, 219, 223, 244, 248
- facsimile 145, 159, 163, 244
- faults (*see also* incorrect forms, mistakes, wrong forms) 120-121, 125-126
- find-spot (*see also* location, provenance, site) 21-23, 26-27, 73, 97, 99, 147
- font 79-80, 159, 173-174
- format 24, 51, 74-76, 95, 109-110, 114, 119, 121, 141, 147, 151-152, 167, 173-174, 176-177, 179, 199, 203, 205, 209-210, 219, 223, 225, 236, 245-246, 253, 257
- formulae (*see also* sentence formulars) 197-198
- formulaic content 71
- formulaic context 15
- formularly patterns 10
- fragment 4, 47, 52, 122, 167-169, 172, 177, 178
- fragmentarily attested languages (*see also* under-resourced languages) XVIII, 14



- gazetteer XV, XVII, 13, 142, 152, 163, 168, 197, 202-204, 206, 208-211, 213, 249, 254-256  
 genealogy 7, 12, 107-108, 110, 112, 114, 207, 235  
 genre (*see also* textual typology) 5, 61, 118-121, 144, 148, 168, 235  
 geography 182, 205  
 geographical area 13, 97  
     geographic(al) distribution (*see also* cartographic distribution, spatial distribution) 11, 31, 183, 236  
     geographical information (*see also* spatial information) XVI, 27, 152  
 geographic(al) name (*see also* names of places, place name, toponym) 53, 74, 163, 256  
 geo(-)referencing XVI, 44, 141, 147, 150-152, 157  
 glossary (*see also* lexicon) XVI, 50, 128, 142, 144-146, 150-152, 155, 162  
 glyph 52-53, 69, 72, 79-80, 102, 107-108, 110-112, 114-115  
 god names (*see also* divine name, names of gods, theonym) 38  
 GPS 27, 106  
 graffiti XV, 9, 11, 39, 104, 106, 115, 187, 209, 217, 219  
 grammar 22, 62-63, 81, 93, 108, 110, 112, 126, 136-137, 254-255  
     grammatical category 52, 123  
     grammatical feature 11, 107, 134-135, 162  
 granularity 11, 50  
 graph 21-25, 28-31, 33, 66-67, 69, 71-73, 75-80, 197, 199, 245, 253-254  
 grapheme XVIII, 21, 23, 29, 31, 38, 42-43, 45, 53, 67, 71, 76, 124-125  
     graphemic sequence 53, 55, 59  
     graphemics 22-24, 72  
 graphic material (*see also* image, visual documentation) 98, 100  
 graphotactic strategies / graphotactics 71, 80  
 guidelines 8, 212, 220, 223, 255  
 hackathons 235  
 harmonization (*see also* alignment, mapping, reconciliation) XVII, 1, 3, 10, 217, 219, 223, 255  
 harvest 12, 216, 233, 235-236, 255  
     harvester 236  
 heterograph (*see also* variant forms) XVI, 118, 125  
 hieroglyphs 67, 71, 73, 77, 80-81, 155, 157-158, 174  
     hieroglyphic text/inscription XVI, 54, 65, 71, 72, 76, 79-81, 155, 157, 159, 161, 163  
     hieroglyphic writing 65, 67, 69, 79-81  
 history of art (*see also* art history) XIII, 6, 181, 254  
 homograph XVI, 118, 123-124  
 homophony 53  
 host 88, 171  
     hosting 91, 110, 130, 142, 147, 157, 169, 171, 175, 180, 227, 228, 231, 243, 254, 257  
 iconography 80, 181, 187, 237, 254  
     iconographic apparatus 256  
     iconographic context 157  
     iconographic elements 6, 26  
 ID / identifier XVI, 4, 32, 47, 88, 91, 110, 147, 151, 158, 193, 197, 199, 203, 206-211, 217, 219, 223-225, 243, 255  
 image (*see also* graphic material, visual documentation) XV, 5, 7, 27, 28, 33, 43, 75, 80, 98, 106, 110, 112, 169, 189, 199, 203, 217, 232-233, 237, 244-245, 253, 256  
     image based search 217  
 import (*see also* capture, incorporation) 88-89, 187, 243  
 incorporation (*see also* capture, import) 242-243  
 incorrect forms (*see also* faults, mistakes, wrong forms) XVI, 118, 123, 125, 127  
 incorrect readings (*see also* misreading) 121  
 index XVII, 6, 53, 58, 155, 161, 168-169, 194, 232, 236, 255  
     indexing 111, 161-162, 172, 180, 224, 232, 247, 255  
 infixes 136  
 infixation 69  
 instance (*see also* attestation, occurrence) 39, 54, 56, 58-60, 119, 253  
 integrations (*see also* restorations) 253  
 intellectual property (*see also* copyright, rights) XVI  
 interface XVI, 5, 32-33, 60, 141-142, 144, 146-147, 150, 152, 159, 162-163, 172, 178, 194, 206, 208, 211, 233, 235-236, 242-244, 247  
 interoperability XIV-XV, XVII-XVIII, 1, 3, 9, 12-13, 44, 82, 87, 89, 160, 162, 165, 193, 198-199, 202-203, 213, 231, 236-237, 245, 248, 253-256

- interpretation XV, XVIII, 8, 15, 21, 26-29, 31, 33, 35, 52-54, 56, 59-60, 71-73, 107-108, 119, 128, 212, 253-254
- JSON 152, 209-210
- KML 236
- lacunae XVII, 7, 134, 140, 253
- laser scanning 203
- lemma / lemmata 58-59, 122, 124-125, 126, 127, 129-130, 136, 148, 254
- lemmatization 141, 144
- lemmatizer XVI, 150-151, 235
- lemmatizing XVI, 141-143, 145-146, 149-152, 198, 235
- lexeme 55, 118-119, 121-130, 171, 254
- lexicon / lexica (*see also* glossary) XV, XVI, XVIII, 14-15, 42, 44, 71, 118, 128, 162, 187-188, 247, 255
- lexical analysis 157, 161
- lexical category (*see also* PoS) 122
- lexical element/entity/entries/item 10, 15, 44, 50, 58-60, 115-116, 121, 255
- lexicography XV, 1, 3, 13, 42-43, 65, 102, 118, 123, 235
- license 152, 163, 237
- licensing 110
- ligature 53, 69
- linguistics XIII, 13, 67, 180, 182, 198, 231, 235, 241
- linguistic analysis 4, 73, 80-81, 186, 188-189
- linguistic unit 55-56
- list of onomastic elements (*see also* onomastic indexes) 106, 112
- literacy 22, 39, 102, 190
- literary texts XV, 1, 168
- literature (*see also* bibliography) 26, 71, 73-75, 119-120, 123, 125, 128-129, 141, 168, 172, 177, 185, 232, 235
- location (*see also* find-spot, provenance, site) 13, 26-27, 44, 72, 80, 106, 114, 159, 171, 177, 186, 188, 232, 236, 242, 245
- Linked Data / Linked Open Data / LOD XVI, XVII, 139, 199, 202, 204, 206-207, 212, 227, 237, 240, 242-244, 248-249, 255
- logogram 53-54, 58
- logographic writing systems XV
- logo-syllabic writing systems XV, 49, 52, 54, 56, 61, 67, 253
- maintenance XVI, 109, 116, 171, 174-175, 179, 218, 245
- maintain 106, 172, 186, 218, 222, 232, 241-244, 249, 257
- manuscript XIII, XV, 8, 22, 86-87, 90-92, 120, 247, 254
- map 40-41, 43-45, 66, 85, 97, 142, 147-148, 150, 152, 213, 233, 244
- map interface / map-based interface XVI, 141-142, 147, 152
- mapping XVII, 11, 12-13, 31, 33, 57, 75, 216, 219, 221, 223, 238, 242, 249, 255-256
- mark-up / markup (*see also* annotation, encoding, tagging) 4, 10, 72, 80, 88-89, 111, 176, 221-222, 226, 233, 245-247
- marking-up 79, 109, 246
- mark-up language 49, 51-52, 54, 57, 59, 62
- match 10, 62, 207, 210-211
- matching 10-11, 50, 52, 59, 62, 144, 151, 210-211, 221, 224, 254-255
- material context (*see also* archaeological context) 209
- meaning XVI, 51, 55-57, 69, 72, 78, 86, 118, 122-125, 128, 134, 150, 199, 205-206, 233
- metadata 1, 3-5, 10-11, 51-52, 80, 87, 102, 106-107, 110, 115, 148, 160, 170, 187, 193, 198-199, 203, 208, 232, 236-237, 241, 243-244, 246, 248
- migration 3, 218
- mining (*see also* search, query) 50-51
- misreadings (*see also* incorrect readings) 123, 125
- mistakes (scribal) (*see also* faults, incorrect forms, wrong forms) 53, 123
- modelling (of data) (*see also* conceptual model, data model) XIV-XV, XVII-XVIII, 1, 3, 19, 72, 77-78, 87, 206, 212, 253-254
- monogram 6-7, 86
- monument 12-13, 24-26, 39, 65, 67, 73-74, 156, 158-159, 163, 181, 183, 185, 231, 233, 264
- morphograph 67, 77-78
- morphographic writing system 67
- morphology XVIII, 40, 69, 129, 187-188
- morphological ambiguity 14, 254
- morphological analysis XVI, XVIII, 95, 118, 122, 157, 253-254
- morphological attributes XVI, 133, 255
- morphological features 89, 162
- morphological tags 122, 129
- morpho-syllabic spellings/writing systems 67, 69, 80-81

- multi-level annotation/annotation system  
49-50, 56, 73, 81
- multilingualism 15
- multilingual inscriptions (*see also* bilingual inscriptions, trilingual inscription) 90
- multilingual textual corpora 256
- multilingual thesauri (*see also* bilingual terminology) 74
- multiple editions 225
- museum 3, 6, 28, 98, 109, 116, 145, 159, 168, 186, 206, 240, 242, 244-245, 248-251, 256
- name (*see also* onomastics) 6, 11, 13, 26, 54-55, 62, 77-78, 85, 107-108, 112, 118-119, 122, 130, 134, 142, 144-146, 151, 161, 194, 197, 198, 204, 235-237, 243, 247
- names of gods (*see also* divine name, god names, theonym) 70, 129
- names of individuals (*see also* anthroponyms, personal name) 12
- names of places (*see also* geographical name, place name, toponym) 12
- named entity 89, 204
- namesakes 150, 151
- NER / Named Entity Recognition 196-197, 255
- network analysis 197, 256
- NLP XVI, 254
- non-assimilated forms (*see also* assimilation) 121
- normalization 97, 151
- normalise 97, 242, 243
- notation 58-59, 67
- OAI-PMH 12
- object-oriented language 59
- object-relational mapping system 57
- occurrence (*see also* attestation, co-occurrence, instance) 14, 15, 29, 50, 55, 58-59, 139, 162, 235, 255
- onomastics XVI, 10, 12, 38, 41, 112, 122, 128, 240, 253, 255
- onomastic database 44, 167, 172, 290
- onomastic indexes/lists (*see also* list of onomastic elements) 12, 168-169
- onomasticon / onomastica 108, 116, 178
- ontology XVI, 73-74, 77, 90, 209-210, 212, 231, 237, 249, 255
- Open Access XIV, XIX, 6, 12, 109, 115, 141, 143, 146, 150, 152, 163, 167, 171, 210, 237, 243, 255
- open source 40, 167, 173-174, 245, 247
- ordered hierarchy of content objects 51, 253
- origin (*see also* site) 13, 27, 135, 186, 188, 254
- orthography 81, 122
- orthographic units (*see also* word-phrase) 9
- orthophotographs XVI, 155, 160
- ostraca 194
- overlap 52, 54, 56, 59, 69
- overlapping 4, 49, 51-53, 56, 84, 88, 253
- OWL 209-210
- painting 182, 186-187, 190
- palaeography 160, 170, 180-181
- palaeographical commentary 187
- palimpsests 53
- paper edition (*see also* printed edition) 2, 186, 257
- papyri 194, 196-198, 254
- papyrology XIII, 195, 200-201
- papyrologists 218
- period (*see also* date-range) XVII, 12-13, 22-23, 85-87, 104, 109, 125, 136, 142-146, 148, 167-168, 170, 189-190, 195-198, 202, 204-213, 232, 235, 237, 240-242, 251, 256-257
- periodization 13, 202, 205, 207
- person(al) name (*see also* anthroponyms, names of individuals) 39, 41, 44, 53, 60, 110, 113, 172, 196, 199, 247
- philology / philological study/research XIII, 4, 9, 13, 26, 59, 61, 233, 235
- philological text editions 168
- philologists 61-62, 181, 225
- phoneme 9, 28, 71, 115
- phonemic system 24, 28
- phonemic value 72, 76
- phonetic value XVIII, 42, 45
- phonology 187-188
- photogrammetric techniques 160
- photograph / photos (*see also* picture, orthophotographs) 16, 23, 44, 95, 97-98, 102, 106-107, 112-113, 115, 120, 145, 156, 159-161, 163, 167-169, 172, 174, 177, 187, 189, 253, 256
- photographic documentation/material/reproduction 94, 120, 123, 125, 163, 181, 187
- phrase 9, 112-113, 125, 128
- picture (*see also* photograph) 37, 40, 43, 108, 182, 184, 233
- place 12, 23, 25, 27, 43, 44, 73, 85, 97, 98, 102, 142, 144-145, 150, 152, 161, 172, 186, 190,

- 193-194, 196-197, 199, 203-204, 206, 210, 213, 235-237, 254-256
- place name (*see also* geographical name, names of places, toponym) 12, 75, 110, 113, 150-151, 163, 172, 177, 196, 198, 203-204, 210, 248
- platform 35, 51, 62, 106, 109-110, 119, 142, 147, 150, 168-169, 171, 174, 180, 193, 195, 203, 211-212, 231, 238, 255
- polyphony 53
- portal XVIII, 142, 144-146, 171, 216-224, 257
- PoS / part of speech (*see also* lexical category) 14, 78, 136, 151, 198, 235-254
- prefix 120, 135-136
- printed edition (*see also* paper edition) 62, 187
- propositional logic 71-72, 78, 253
- prosopography 151, 193-194, 196, 215, 263, 289
- provenance (*see also* find-spot, location, site) 13, 23, 27, 106, 108, 111, 114, 168, 180, 196-197, 199, 224, 233, 242, 254
- public (use of) script 180-181, 185, 190
- query (*see also* search) 4, 9, 13, 24, 49-50, 59, 61-62, 73, 209, 225
- queried 89, 177, 244
- querying 50, 61, 224
- quotations 85, 126, 129-130, 232
- RDF 73, 75, 77, 160, 209, 236, 238, 243, 248-249, 253, 255
- reading (*see also* interpretation) XVI, 8, 10, 31, 43, 45, 52-54, 56, 59-60, 62, 71-72, 77-81, 95, 107, 115, 118, 120-121, 123, 127, 134, 157, 253
- reconciliation (*see also* alignment, harmonization) 202, 206-207, 210
- regular expressions 43-44, 50, 62, 175
- relational database/model 4, 9, 28, 33, 40, 49, 52-53, 57-59, 61, 95, 110, 111, 119, 194, 253-254
- repository XVI, 12, 49, 51, 54, 61, 75, 87, 110, 160, 209-210, 243, 254-255
- responsibility 5, 145, 249, 251
- restorations (*see also* integrations) 7-8, 95
- retrieval 2, 4, 8, 17, 54, 172
- retrieve 8, 14, 128-129, 151, 177, 254
- reuse XVII, XVIII, 75, 86, 157, 162, 176, 199, 203, 210, 217, 219, 225, 233, 235-236, 256-257
- rights (*see also* intellectual property, copyright) 75
- root 14-15, 62, 113, 116, 121-122, 125-126, 133, 135-138, 140, 193, 198, 254
- schema XVII, 4, 10, 12, 49, 57, 72-78, 80-81, 88, 91-92, 219-220, 233, 236, 255
- search (*see also* mining, query) XVI, XVII, 4, 8-9, 15, 24, 31, 35, 41-45, 50, 60, 76, 89, 107, 110-112, 126-127, 136, 139, 143-145, 148, 160-162, 172, 202, 204, 206, 208-209, 211, 213, 216, 219, 223-224, 233, 244, 256
- searchable XV, 21, 25, 102, 106, 108-110, 141, 145-146, 175, 178, 185, 217, 253
- searching 58-59, 96, 208, 212, 222, 224, 243
- semantic mark-up/modelling/relations 13, 77, 88, 237, 253
- semantic web 212, 231, 237
- semantics 80, 256
- semantic differences 15, 124
- semantic units 55-56, 130
- semantic value XVIII, 8, 253
- semi-automatic annotation 80-81
- semi-syllabic writing system 38
- sentence formulars (*see also* formulae) 140
- siglum / sigla 98, 110-111, 113, 116
- sign 6, 10, 24-25, 28-29, 31, 34, 43-44, 50, 52-54, 56-58, 60, 62, 66-67, 69, 71-73, 76-81, 148, 152, 157, 163, 253-254
- site (*see also* location, origin, provenance, find-spot) 5, 13, 44, 52, 55, 66, 68, 106, 114, 148, 150, 155-156, 161, 163, 167-168, 186, 203, 236
- SKOS 74-75, 210, 248
- space 12, 28, 203-206, 213
- spatial context/information/relation 6, 13, 72-73, 80, 202, 235, 238
- spatial coverage/extent 206, 211-212
- spatial entity 203-206
- spatial distribution (*see also* cartographic distribution, geographical distribution) 50, 232
- squeezes 106, 133-134, 140
- standard 82, 109, 115, 123, 169, 173-175, 178, 195, 203, 205, 209, 211, 213, 216, 220, 242, 244, 253, 255
- standardization 75, 79, 81, 97-98, 126, 152, 167, 199, 243, 245, 256
- sticks 9, 23, 103, 118
- string 50-51, 56, 58-59, 62, 120, 126, 129, 197, 219, 224, 253

- structure of the text (*see also* text structure) 7, 51, 53, 80, 90
- suffix 120, 126, 129, 135-137
- support (*see also* archaeological object, artefact, carrier, text-bearing object) 5, 6-7, 11-12, 90-91, 148, 180, 225, 233, 253
- sustainability X, 16, 167, 174, 178-179, 186, 257
- syllabary 57
- syllabic graphemes/signs/value 38, 53, 67, 78
- syllabograms 43, 53
- syllabographs 67
- symbol 7, 25-26, 232, 235-237
- syntagmatic units 62
- syntax 40, 67, 140, 187-188
  - syntactic annotation 177
  - syntactic features 10, 67
- tablet 39, 49, 52, 54-55, 58-60, 167-170, 172-173, 177-178
- tag 4, 11, 112, 122, 125, 129, 176, 187, 224
  - tagged 12, 107, 112, 122, 134, 140
  - tagging (*see also* annotation, encoding, mark-up) XV, XVI, 7, 9, 79, 98, 109-110, 122-123, 135, 140-141, 144, 176, 198
- taxonomy 183, 255
- teaching 4, 245
- TEI 4, 7-8, 10, 51, 65, 69, 79-81, 87-88, 90, 92, 109, 162, 212, 216, 224, 227, 236-238, 242-243, 247, 253
- terminology XVII, 4, 10-11, 24, 31-32, 34, 75
  - terms XVII, 4, 6, 10-11, 58-59, 74-75, 102, 108, 136-137, 161-163, 182-183, 205-206, 208-212, 219-220, 232, 247, 249
- tesserae 39
- text-bearing object (*see also* archaeological object, artefact, carrier, support) 4, 66, 72-73, 75, 233, 253
- text direction (*see also* direction of script) 90
- text edition 87, 119, 128, 141, 147, 159, 168, 170-172, 176-178
- text structure (*see also* structure of the text) 80, 90
- (textual) context 4, 14, 50, 78, 108, 122, 125, 129, 150, 161-162
- textual criticism 180-181
- text(ual) typology (*see also* genre) 10, 219
- theonym (*see also* divine name, god names, names of gods) 40-41, 44, 122, 125, 161
- thesauri (*see also* taxonomy) 74, 237, 255
- time (*see also* chronology, date, period) 23, 25, 28, 93, 125, 135-136, 182, 190, 202, 204-206, 232, 235, 251, 257
- tokens 60
  - tokenized 50
- toponym (*see also* geographical name, place name) 13, 44, 122, 125, 161-163
- transcription XVI, 4-5, 8, 10, 42-44, 69, 76, 80, 87-88, 90, 97, 120-121, 150, 174, 253
- translation XVI, XVII, 1, 3, 5, 14-15, 26-27, 32, 59, 62, 85, 102, 107, 110, 113, 118-119, 125-130, 133, 136, 138-140, 142-146, 148, 150-152, 202, 217, 219, 222, 243, 256
- transliteration XV, XVI, 4, 8-9, 21, 26-27, 31, 34, 49, 53, 59, 72-73, 76, 78-80, 89, 91, 110, 113-115, 129, 143-146, 148, 152, 169, 174, 178, 197, 253-254
- trilingual inscription 84-85, 143
- triple store 75
- type of text (*see also* textual typology) 195
- typography 80
  - typographic rendering 56
- uncertain reading 8, 10, 123
- under-resourced languages (*see also* fragmentarily attested languages) XIV, XVI, 1, 13, 254
- Unicode 9, 43, 50, 79, 174, 251
- up-conversion XVII, 216, 219, 221-223, 255
- URI 73, 79, 80, 158, 199, 209-210, 212, 242, 244, 247-248
- URL 75, 112, 114, 142, 210
- variant / variation XVII, 10, 38-39, 42-43, 253-254
  - glyph/graph(ic)/graph-type/script variant/variation (*see also* allographic notation) XV, 21, 23-25, 27-31, 33, 53, 72, 77, 79, 102, 115, 157, 163, 253
  - (reading) variant/variation XVI, 8-9, 43, 45, 56, 60, 118, 120-121
  - variant (forms)/variation (*see also* heterograph) 116, 125, 128
- version 44, 90, 171, 243
  - versioning XVI
- virtual research environment 65, 73, 75, 77, 82, 87
- visual analytics 233
- visual documentation (*see also* graphic material, image) XV, 3-7, 21, 253
- visualisation 201, 231
  - visualising 31, 197, 236, 239

- vocabulary (*see also* controlled terms, thesauri) 10, 73-75, 121, 206-209, 217-220, 223, 226, 237, 242, 248, 255
- web-GIS 50
- word form 14, 177, 232, 235
- word patterns (*see also* regular expressions) 9
- word-phrase (*see also* orthographic units) 133-137, 139-140
- workflow 4, 75, 128, 130, 219, 223
- writing surface 185-186, 195, 199
- wrong forms (*see also* faults, incorrect forms, mistakes) 123
- XML XV, 1, 3-5, 7-9, 11, 49, 51-52, 54-55, 61, 65, 69, 79-81, 87-88, 90-92, 108-109, 111, 119, 121-122, 174-177, 179, 187, 198, 202, 210, 212, 217, 219-220, 222-223, 236-238, 242-247, 250, 253-254