Archeologia e Calcolatori
34.1, 2023, 157-164
doi 10.19282/ac.34.1.2023.17

CHALLENGES IN RESEARCH COMMUNITY BUILDING: INTEGRATING TERRA SIGILLATA (SAMIAN) RESEARCH INTO THE WIKIDATA COMMUNITY

1. Introduction

Community building creates a network. This also applies to data which may result in a knowledge graph. In the case of graph-based models, nodes and edges can be modelled in the Resource Description Framework (RDF) standard based on a semantic network as Linked Open Data (LOD) being a part of the Semantic Web. In the RDF model, each statement consists of the three units subject, predicate and object, whereby a resource as a subject is described in more detail by another resource or a value (literal) as an object. With another resource as a predicate, these three units form a triple ("3-tuple"):

(Subject) -[Predicate]-> (Object)

As an example, the ancient Greek philosopher Plato is a human entity, that can be described in RDF using common, well-known Semantic Web vocabularies and ontologies, especially the RDF Schema (Brickley, Guha 2014) and the Friend of a Friend (FOAF) vocabulary (Brickley, Miller 2004):

(ex:Plato) -[rdf:type]-> (foaf:Person)

The four LOD principles (Berners-Lee 2006) should be applied to domain-specific archaeological data to create an Archaeological LOD Cloud as part of the Giant Global Graph, the Linked Open Data Cloud (more about LOD in archaeology can be studied by Isaksen 2011; Thiery 2013; Schmidt *et al.* 2022). One possibility to create a direct link into the LOD Cloud and integrate volunteers and citizen scientists is Wikidata (Vrandečić, Krötzsch 2014). Wikidata is a free and open knowledge base, a secondary database as well as a data hub where everybody can add and edit new entities and classes. Wikidata is the central storage for structured data of projects by the Wikimedia Foundation, such as Wikipedia and Wikimedia Commons. Data in Wikidata is available under a free licence (CC 0), multilingual, accessible to humans and machines (GUI, API, SPARQL), exportable using standard formats (JSON, RDF, XML) and interlinked to other open data sets in the LOD Cloud. The English author and screenwriter Douglas Adams, best known for *The Hitchhiker's Guide to the Galaxy* where "42" is the answer to the ultimate question of life, the universe, and everything, can be described in Wikidata as:

(wd:Q42) -[wdt:P31]-> (wd:5)

This means Wikidata Entity Douglas Adams (Q42, English science fiction writer and humorist) is an instance of (P31, a particular example and

member) a human (Q5, the common name of *homo sapiens*). Q42 is commonly used to describe the Wikidata data model (Fig. 1). Wikidata's data model consists of identifiers, labels, descriptions, and aliases, as well as statements such as properties, values, qualifiers, and references. E.g., Plato is described with the identifier Q859 as an instance of human, with male gender, with classical Athens citizenship, date of birth 347 BCE (Gregorian), working in the fields of philosophy, literature and politics.

Wikidata properties and items suffice for a large range of Roman ceramics data, reflecting a diverse and active community of users but also diverse implementations of data models. We would like to discuss the benefits and challenges of integrating communities. The Leibniz-Zentrum für Archäologie, Mainz (LEIZA) curates the Samian Research database, a treasure-house of economic data on Roman trade and the Terra Sigillata (Samian Ware) industry. Over six decades, a broad European user community of established research institutions, citizen scientists and domain-specific scientists has assembled a dataset of ca. 250,000 potter's stamps from the Samian Research database, accessible with findspots and relevant bibliography as Linked Open Samian Ware (LOSW) via the collaborative LOD hub "archaeology.link".

For this purpose, a reproducible workflow (Fig. 2; Thiery *et al.* 2021) was developed to transform the Samian Ware data from its original relational structure into LOD and FAIR data (Wilkinson *et al.* 2016) to reuse the data. First, entries such as potter stamps are curated in an interactive web application

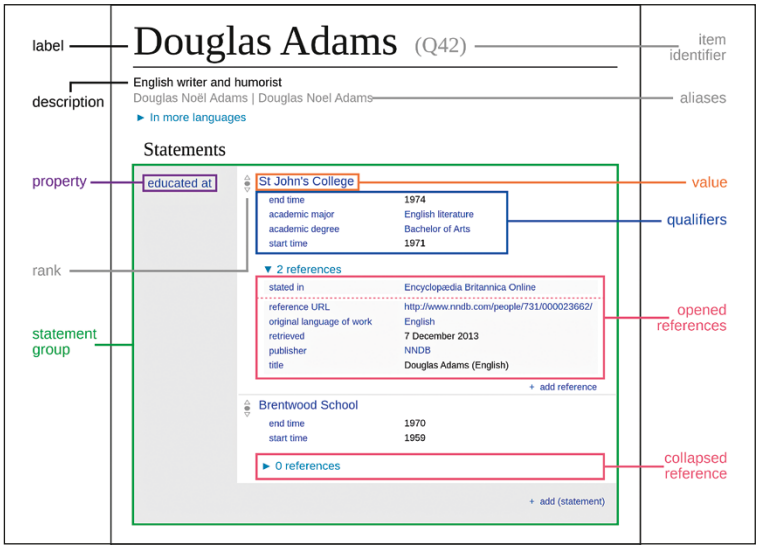

Fig. 1 – Graphic representing the data model in Wikidata with a statement group and opened references (C. Kritschmar, WMDE, via Wikimedia Commons).

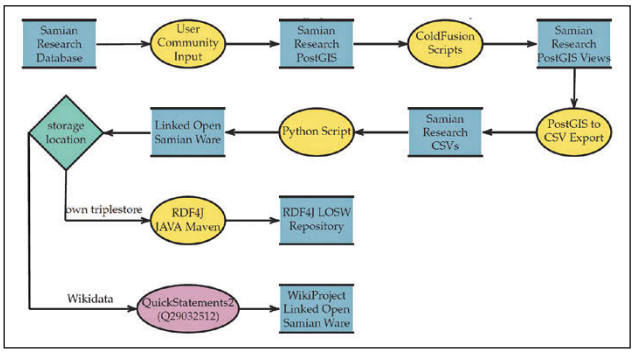

Fig. 2 – Linked Pipe: Linked Open Samian Ware, as data flow diagram using the Yourdon And/Or De Marco notation and the Linked Pipes style (via Thiery *et al.* 2021).

Samian Research and stored in a PostgreSQL database. Second, these data are exported to CSV files, which are then transformed to RDF using Python scripts according to the Samian Ontology (Thiery, Mees 2021). To share the data, two sub-workflows are implemented: (i) one leading to a self-hosted triple store, where the data is mapped to places in the Pleiades gazetteer and to Roman ceramic typologies in the Ceramic Typologies Ontology (CeraTyOnt) (Thiery *et al.* 2020), (ii) the other one to the Wikidata WikiProject "Linked Open Samian Ware" (Thiery, Mees 2022) using "QuickStatements" to transform the CSV data to Wikidata entries (Schmidt *et al.* 2022).

2. Samian Research in Wikidata

In 2020-2021, Samian Research began a process of integrating its data within Wikidata through the creation of a set of Samian Ware Wikidata items, including 3,874 Samian Ware Discovery Sites, 103 Samian Ware kiln sites and 13 kiln regions, comprising accurate or approximate geospatial information and a backlink to the LOD Hub "archaeology.link". Within Wikidata, geospatial classes describing Samian Ware data were created: Samian Ware Discovery Sites (Q102202066), production centres (kiln sites) as Samian Ware Production Centres (Q102202026), and kiln regions as Samian Ware Kilnregion (Q102201947). Each geospatial resource is also categorised as an "archaeological site" (Q839954) (Schmidt *et al.* 2022).

These three new classes are characterised by several unique attributes, identified as derived from the Samian Research database. Samian Ware Discovery Sites in Wikidata (Fig. 3) are instances of (P31) Samian Ware Discovery Site (Q102202066), are part of (P361) Samian Research (Q90412636) and have exact matches (P2888) as backlink to the Linked Open Samian Ware

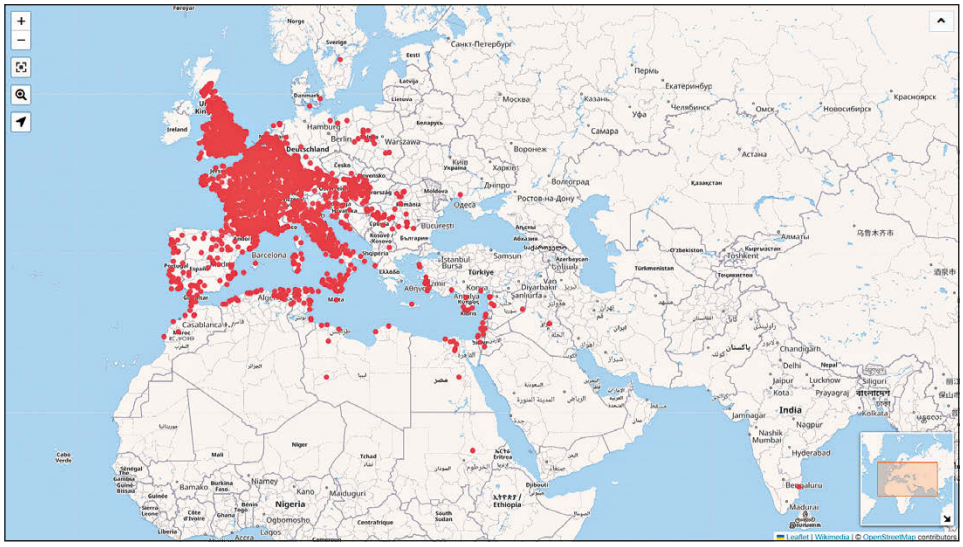

Fig. 3 – Linked Open Samian Ware Discovery Sites (red dots) on Wikidata queried via https://w.wiki/6EBu on 09/01/2023, Wikidata Community.

Fig. 4 – Linked Open Samian Ware Production Centres (red dots) on Wikidata queried via https://w.wiki/6EBx on 09/01/2023, Wikidata Community

URI, e.g. the Pompeii Samian Ware Discovery Site (Q103190089) with its URI http://data.archaeology.link/data/samian/loc_ds_1003977 (located at 40°45'00.0"N 14°28'60.0"E).

Samian Ware production centres in Wikidata (Fig. 4) are instances of (P31) a Samian Ware Production Centre (Q102202026). They are also part of (P361) Samian Research (Q90412636) and have exact matches (P2888) as backlink to the Linked Open Samian Ware URI, e.g. the La Graufesenque Samian Ware Production Centre (Q102763431) with its URI http://data.archaeology.link/data/samian/loc_pc_2000001 (located at 44°06'00.0"N 3°05'00.0"E). Samian Ware kiln regions in Wikidata (Fig. 5) are instances of (P31) a Samian Ware kiln region (Q102201947), are also part of (P361) Samian Research (Q90412636) and have exact matches (P2888) as backlink to the Linked Open Samian Ware URI. These kiln regions are calculated as a convex hull of production centres having the same regional tradition. The production centre La Graufesenque is part of the South Gaulish kiln region (http://data.archaeology.link/data/samian/loc_kr_131462; Q102764958) located in the S of modern France. In the Wikimedia Universe, this geospatial data is also stored in Wikimedia Commons as GeoShape GeoJSON, e.g., the before mentioned kiln region.

Creating designated Wikidata items is an efficient way to map the substantial geographic reach of our subject. It refers to many European archaeological sites and excavations which hitherto lacked a Wikidata identifier. For example,

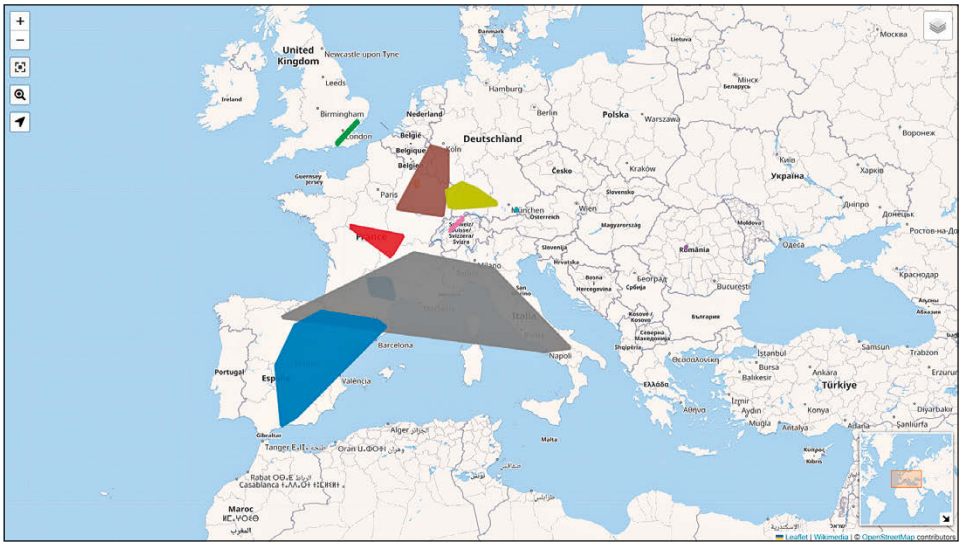

Fig. 5 – Linked Open Samian Ware kiln regions (coloured polygons) are calculated as a convex hull of production centres having the same regional tradition, on Wikidata queried via https://w.wiki/4pDu on 09/01/2023, Wikidata Community.

the Samian Ware Discovery Site "Saint-Georges-de-Reneins". A new findspot in Samian Research from June 2022 has (currently) no "archaeological site" entity in Wikidata. This one-directional way, due to not directly declaring Wikidata entities in Samian Research, does not currently allow its sites to be merged automatically with existing Wikidata items for archaeological sites and excavations.

3. Use case: Corinth

The archaeological site of Corinth illustrates one obvious issue that has to be solved, e.g. the Corinth from Geonames with ID 259289 (https://www.geonames.org/259289/korinthos.html) is not the Corinth of Topos-Text Place ID 379229BEuB (https://topostext.org/place/379229BEuB) and not the Samian Ware Discovery Site Corinth with ID 1003935 (http://data.archaeology.link/data/samian/loc_ds_1003935). But how to create semantic clarity related to archaeological sites, also within Wikidata? Here, more than four overlapping concepts associated with very different Wikidata properties exist; 15 monuments in ancient Corinth have Wikidata items of their own, e.g., theatre, Asklepieion, the temple of Apollo, etc. The following entities are semantically and content related, but not the same:

– Corinth (Q1363688 - the ancient city in historical literature),
– ancient Corinth (Q22681231 - the modern village),
– ancient Corinth (Q101834062 - the organised archaeological site),
– the Corinth excavations (Q5170664 - archaeological excavation, an organised activity).

Semantic clarity can be created by highlighting four main items: excavations, sites, ancient and modern places:

– The excavations are the source of scientific information via scholars, publications, archives, and storerooms.
– The site is something one might visit, under specific circumstances, to see specific things.
– The ancient place has a vaguely defined territory but is associated with specific historical events and real or mythological people.
– The modern place, in Greece, usually re-baptised with an ancient name, has different properties: administrative, population, etc.

These main items cause challenges:

– Wikipedia editors like to combine – unlike Wikidata items – ancient and modern places and sites with their excavations.
– Wikidata editors create double and triple entities for maybe the same thing because they are a bit different. These have to be merged, e.g., Q103160025 as the created entity for the original Samian Ware Discovery Site Corinth (ID 1003935), which was merged into the archaeological site of Corinth (Q101834062).

– Good Linked Open Data keeps different entities separate, using properties to link them in a human and machine-readable way.
– Complete rigour is massively labour-intensive. We combine where two items would each have too little structured data to justify creating it.
– Rescue excavation in Athens is both an archaeological site (a street address) and a short-lived excavation producing limited archaeological data.

Currently, Wikidata entries (e.g., Pompeii - Q43332) combine the ancient city, the modern archaeological site, and excavations. The Wikidata item Q103190089 Pompeii (Samian Ware Discovery Site) does not link to Pompeii Q43332, the ancient city/ruins, except through a shared Pleiades ID and similar coordinates. Nor does Q43332 link to Q103190089. The future praxis should ideally be that "Samian Ware Discovery Sites" are linked via Wikidata to the excavation, which documents the presence of this ware and/or to the archaeological site. We find 2,916 Wikidata items for archaeological sites in Italy but only 5 archaeological excavations. The latter concept is used (for now) primarily in Greece. Meanwhile, ca. 300 Samian Ware site items for Italy. Most could be combined with other Wikidata items if we choose.

4. Conclusion and outlook

To solve these issues, the broader Wikidata Community must be enlisted. Wikidata properties and items suffice for almost the full range of Roman ceramics data, reflecting the diverse and active communities of users but also diverse implementations of data models. Knowledge exchange must be enabled, e.g., by bidirectional links using properties in Wikidata. We currently use P2888 (exact match) which causes problems with multiple assignments. A solution to this can be creating an "archaeology.link property", allowing for multiple exact matches, which remains to be discussed within the Wikidata community. However, in a specialised domain, can community-validated data entries safely generate new knowledge? We think it does, because it may result in information from researchers familiar with the local situation, from which specific excavation in Corinth a particular Samian research object may originate and to add this information to the archaeological Linked Open Data Cloud. We consider Wikidata Community projects in archaeology (Trognitz *et al.* 2023) as an umbrella for community initiatives, e.g., Linked Open Samian Ware to address issues of sustainability and data consistency.

Florian Thiery, Allard W. Mees
Leibniz-Zentrum für Archäologie (LEIZA)
florian.thiery@leiza.de, allard.mees@leiza.de

John Brady Kiesling
ToposText.org
topostext@gmail.com

REFERENCES

Berners-Lee T. 2006, *Linked Data - Design Issues* (https://www.w3.org/DesignIssues/LinkedData.html).

Brickley D., Guha R.V. 2014, *RDF Schema 1.1. W3C Recommendation 25 February 2014* (https://www.w3.org/TR/rdf-schema/).

Brickley D., Miller L. 2004, *FOAF Vocabulary Specification. Namespace Document 1 May 2004* (http://xmlns.com/foaf/0.1/).

Isaksen L. 2011, *Archaeology and the Semantic Web*, Thesis (Doctoral), University of Southampton, School of Electronics and Computer Science (https://eprints.soton.ac.uk/id/eprint/206421).

Schmidt S.C., Thiery F., Trognitz M. 2022, *Practices of Linked Open Data in archaeology and their realisation in Wikidata*, «Digital», 2, 3, 333-364 (https://doi.org/10.3390/digital2030019).

Thiery F. 2013, *Semantic web und linked data: Generierung von Interoperabilität in archäologischen Fachdaten am Beispiel römischer Töpferstempel*, «Squirrel Papers», 4, 1, 3, Thesis (Master), Mainz, Germany, Fachhochschule Mainz (https://doi.org/10.5281/zenodo.292979).

Thiery F., Homburg T., Trognitz M. 2021, *Linked pipe: Linked open Samian Ware*, «Squirrel Papers», 4, 2, 3 (https://doi.org/10.5281/zenodo.5779053).

Thiery F., Mees A.W. 2021, *Ceramic typologies ontology (CeraTyOnt),* «Squirrel Papers», 3, 3, 1 (https://doi.org/10.5281/zenodo.5767082).

Thiery F., Mees A.W. 2022, *Wikidata: WikiProject Linked Open Samian Ware* (https://www.wikidata.org/wiki/Wikidata:WikiProject_Linked_Open_Samian_Ware).

Thiery F., Mees A.W., Gottwald D. 2020, *Linked open Samian Ware*, «Squirrel Papers», 2, 2, 1 (https://doi.org/10.5281/zenodo.5767082).

Trognitz M., Epìdosis, Schmidt S.C., PKM, Anitnelav V., Thiery F. 2023, *Wikidata: WikiProject Archaeology* (https://www.wikidata.org/wiki/Wikidata:WikiProject_Archaeology).

Vrandečić D., Krötzsch M. 2014, *Wikidata: A free collaborative knowledgebase*, «Communications of the ACM», 57, 10, 78-85 (https://doi.org/10.1145/2629489).

Wilkinson M.D., Dumontier M., Aalbersberg Ij.J., Appleton G., Axton M., Baak A., Blomberg N., Boiten J.-W., da Silva Santos L.B., Bourne P.E., Bouwman J., Brookes A.J., Clark T., Crosas M., Dillo I., Dumon O., Edmunds S., Evelo C.T., Finkers R., Gonzalez-Beltran A., Gray A.J.G., Groth P., Goble C., Grethe J.S., Heringa J., 't Hoen P.A.C., Hooft R., Kuhn T., Kok R., Kok J., Lusher S.J., Martone M.E., Mons A., Packer A.L., Persson B., Rocca-Serra P., Roos M., van Schaik R., Sansone S.-A., Schultes E., Sengstag T., Slater T., Strawn G., Swertz M.A., Thompson M., van der Lei J., van Mulligen E., Velterop J., Waagmeester A., Wittenburg P., Wolstencroft K., Zhao J., Mons B. 2016, *The FAIR Guiding Principles for scientific data management and stewardship*, «Scientific Data», 3, 160018 (https://doi.org/10.1038/sdata.2016.18).

ABSTRACT

In 2020, the Samian Research database began a process of integrating its data within Wikidata through the creation of a set of Samian Research Wikidata items, including Samian Ware Discovery Sites, Samian Ware kiln sites and kiln regions, comprising accurate or approximate geospatial information and a backlink to the Linked Open Data hub 'archaeology.link'. This approach of creating designated Wikidata items is an efficient way to map the enormous geographic reach of our subject and to call attention to many European archaeological sites and excavations that hitherto lacked a Wikidata identifier. The site of Corinth illustrates an exemplary issue to be solved: ambiguity and different archaeological concepts and ideas. E.g., is it correct to merge Corinth as a Samian Ware Discovery Site with the archaeological site of ancient Corinth? To solve the issue, the broader Wikidata community must be enlisted. This paper describes the challenges in the use case of Corinth and offers solutions within Wikidata.