Archeologia e Calcolatori
34.1, 2023, 15-20
doi 10.19282/ac.34.1.2023.02

THE HARRIS MATRIX DATA PACKAGE SPECIFICATION AND THE NEW INIT COMMAND OF THE PYTHON HMDP TOOL

1. Harris Matrix Data Package

Harris Matrix Data Package is a lightweight and user-oriented format for publishing and consuming archaeological stratigraphy data. Harris Matrix data packages are made of simple and universal components, they can be produced from ordinary spreadsheet or database software and used in any environment. Harris Matrix Data Package is entirely based on the data tables proposed for the "hm" Lisp package (Dye, Buck 2015), with the important addition of metadata from the Frictionless Standards specifications. As the "hm" documentation states «There are seven data tables potentially used as input to hm. A project requires a table for units of stratification and for observed stratigraphic relations, but the other five tables are optional». Data tables are CSV files and the metadata descriptor is a JSON file.

Harris Matrix Data Package was first presented at ArcheoFOSS 2019 (Costa 2019), but it has not seen any significant adoption, despite important work in the area of long term and sustainable archiving of stratigraphy data (Moody *et al.* 2021). When looking at the limitations of the initial proposal for the data format, the lack of a clearly documented specification, that must be kept separate from the software tool, was identified as the main issue. The data format specification is now presented as an RFC-style document at the stable URL https://www.iosa.it/specs/harris-matrix-data-package/.

From the point of view of Frictionless Standards, Harris Matrix Data Package is a Data Package "profile" similar to the Fiscal Data Package (Walsh *et al.* 2018) or the Camera Trap Data Package (Camtrap DP Development Team 2022). The following definitions apply in the context of the Harris Matrix Data Package specification:

– "data descriptor" is a JSON file, named "datapackage.json", that is found in the top-level directory of a data package, and contains metadata about the entire data package (name, description, creation date, author names, references) together with the data package schema;
– each "resource" is a CSV table;
– "contexts" refer to archaeological stratigraphy units as produced by the single context recording method; contexts can be both positive and negative and are described in terms of unit-type and position;
– "observations" refer to the stratigraphic relationship between pairs of contexts, and can only record relative chronology of earlier-later relationships

– sameness and contemporaneity of contexts are treated separately in the "inferences" table;
– "inferences" refer to once-equal contexts that were recorded separately, but get treated as a whole for the purpose of stratigraphy, as is the case of a floor level that was divided in two separate units by a later trench;
– "phases" and "periods" are groupings of contexts that are based on chronological affinity;
– "events" are associations between absolute chronology events and contexts, and the resource specifies the nature of the association (Dean 1978);
– "event-order" is an indication of the order of absolute chronology events related to the same context.

The simplest Harris Matrix Data Package will contain three files in one directory: datapackage.json, contexts.csv and observations.csv. This basic layout is capable of recording all directly observed stratigraphic relationships, that can be processed into a directed acyclic graph (DAG) and visualized as a Harris Matrix. The five optional tables alter the graph, by providing instructions for merging one or more stratigraphic units in a single graph node, for grouping nodes with visual indication of their phase or period, for adding nodes to the graph from absolute chronology events, in a specific order. At present, the format can be read using the "hm" Lisp software, the "hmdp" Python software described below and the R "stratigraphr" library (with a few intermediate steps through the "frictionless" library). The only known implementation of a writer is the "hmdp init" command.

The Harris Matrix Data Package is not necessarily the best native format for an archaeological information system, because they are typically modeled as relational databases with a much wider scope. However, the combination of simplicity in the choice of file formats with detailed machine-readable metadata is a good tradeoff that should make it a good, if not optimal, exchange and archival format.

2. The hmdp tool

The hmdp tool (always spelled in lower case) is a command line program with three separate subcommands: init, check and matrix. This tool is developed as a minimalist and versatile proof of concept for a more complete stratigraphy software package, but it could also be reused as part of other existing applications, like pyArchInit (Mandolesi, Cocca 2013). The version that is referred to in this paper is hmdp 2022.10.16, released under the GNU GPLv3 license (Costa 2022). hmdp is written in the Python programming language and depends on foundational open source libraries such as NetworkX, pygraphviz and Graphviz (Gansner, North 2000; Hagberg, Schult, Swart 2008) for processing of stratigraphy data into a directed acyclic graph. The data package is read into the program with the functions provided by the "datapackage" and "goodtables" Python libraries (now superseded by the "frictionless" library).

2.1 *The new init command*

The init command can create a new, empty data package from scratch, consisting of the metadata descriptor and empty data files, easing the creation of data packages that otherwise would require several manual steps and editing of text files. The naming and functionality are inspired by the well-known "git init" command, although in this case "hmdp init" will accept command line arguments for the data package name, author, and resource presets (described below in further detail). If no argument is provided, an interactive prompt in the terminal asks the user to provide the same data and populates the datapackage.json file.

Based on the amount of information available, "hmdp init" provides four presets with an increasing level of detail (the numbering in the list correspond to the numbers associated with each preset in the command line interface):

1. contexts and observations (basic data model);
2. contexts, observations and inferences, to add data about sameness or contemporaneity of stratigraphic contexts, which is necessary to reconcile the difference between the traditional Harris Matrix that allows to define two contexts as equal but still separate, and the need to merge such units in the DAG model;
3. contexts, observations, inferences, periods and phases, a second step towards complexity, with the periods and phases as loosely defined as possible in order to allow multiple levels of chronological periodisation;
4. contexts, observations, inferences, periods, phases, events and event order, the entire set of resources that is only supported by "hm" for analysis but can still be added and managed.

The presets listed above were chosen to simplify the creation of data packages, without forcing users to make all-or-nothing choices. In particular presets 2 and 3 allow to create data packages of intermediate complexity. It is always possible to edit the data package at a later stage, add or remove resources, but this requires small changes to the schema of foreign keys that is not supported by the version of the program described here: for example, if the data package is created with preset 3, then the "contexts" resource schema will contain a "foreignKey" referring the "periods" and "phases" resources, and removing those resources without removing the corresponding "foreignKey" returns a broken data package. A simplistic workaround to all these small limitations would be to always create a complete data package, with example data in all resources including those that are not actually used, but this was not deemed correct. In general, the hmdp init command is agnostic to the subsequent steps taken either with other software or with hmdp itself. The stated aim is, again, to enable different tools to interoperate with the same data format.

2.2 *The check command*

The check command is a useful shortcut to run all available validation tasks on the data package. The command must be given only one argument, the datapackage.json, and will perform three checks on the dataset:

– validate the metadata descriptor without looking at the data (e.g. resources can be missing or broken but the JSON file is well formatted);
– validate every resource for internal consistency (e.g. there are column headers, each row has the right number of columns, constraints like integer values, enums, etc. are respected);
– check the consistency of foreign keys based on the data descriptor.

Ideally, this command would be run after each modification to the resource files and to the metadata descriptor.

2.3 *The matrix command*

The core functionality of hmdp is in the matrix command. The matrix command takes two arguments, the input data package and the output Graphviz DOT file, in the same vein as "hm", even though hmdp is quite limited if compared to its Lisp predecessor. All contexts are read from the "contexts" resource, optionally assigning an attribute for a different graphical shape corresponding to the unit type (the "hm" documentation mentions three possible values: deposit, interface, other). Each context is added to the graph as a node. Observations of all stratigraphic relationships are added to the graph as "directed edges" from the younger node to the older node: this direction of the stratigraphic relationship is consistent with the single context recording methodology. If there are no cycles in the graph and a formally correct DAG can be created, the last processing step is the removal of redundant relationships with the "transitive reduction" function of NetworkX. Transitive reduction is also available as a Graphviz command "tred", but using NetworkX within the same Python program has the advantage of returning a DOT file that does not need further data processing, making the procedure more reproducible and in line with Open Science good practice (Marwick 2017). The graph is exported to the output Graphviz DOT file with the "ortho" layout attribute, so that edges will be drawn as orthogonal polylines.

The resulting DOT file is a digital representation of the Harris matrix that can be visualized in graphical form, with stratigraphic units represented as “nodes”, possibly of different colors and shapes. The most effective way to obtain a static image is to run the Graphviz “dot” command, that will output a PNG or SVG file.

A crucial assumption of this two-step processing (from data to Graphviz, from Graphviz to image) is that multiple visualization outputs are possible from the same initial data. We are moving beyond the idea that the Harris Matrix is an output that is created (either manually or digitally) and then literally looked at in order to obtain information and seek answers, towards a research process where stratigraphy data is transformed and manipulated in a transparent way to obtain different chronological models that help with specific research questions.

Stefano Costa
Soprintendenza ABAP per le province di Imperia e Savona
Ministero della Cultura
stefano.costa@cultura.gov.it

REFERENCES

Camtrap DP Development Team 2022, *Camera Trap Data Package (Camtrap DP)* (https://tdwg.github.io/camtrap-dp).

Costa S. 2019, *Una proposta di standard per l’archiviazione e la condivisione di dati stratigrafici*, in P. Grossi *et al.* (eds.), *ArcheoFOSS. Free, Libre and Open Source Software e Open Format nei processi di ricerca archeologica, Atti del XII Workshop (Roma 2018)*, «Archeologia e Calcolatori», 30, 459-462 (https://doi.org/10.19282/ac.30.2019.29).

Costa S. 2022, *hmdp version 2022.10.16* (https://doi.org/10.5281/zenodo.7213369).

Dean J.S. 1978, *Independent dating in archaeological analysis*, in M.B. Schiffer (ed.), *Advances in Archaeological Method and Theory*, San Diego, Academic Press, 223-255 (https://doi.org/10.1016/B978-0-12-003101-6.50013-5).

Dye T.S., Buck C.E. 2015, *Archaeological sequence diagrams and Bayesian chronological models*, «Journal of Archaeological Science», 63, Suppl. C, 84-93 (https://doi.org/10.1016/j.jas.2015.08.008).

Gansner E.R., North S.C. 2000, *An open graph visualization system and its applications to software engineering*, «Software - Practice and Experience», 30, 11, 1203-1233.

Hagberg A.A., Schult D.A., Swart P.J. 2008, *Exploring network structure, dynamics, and function using NetworkX*, in G. Varoquaux, T. Vaught, J. Millman (eds.), *Proceedings of the 7th Python in Science Conference - SciPy (Pasadena, CA 2008)*, Pasadena, 11-15 (http://conference.scipy.org/proceedings/SciPy2008/paper_2/).

Mandolesi L., Cocca E. 2013, *PyArchInit: gli sviluppi dopo ArcheoFoss 2009*, in M. Serlorenzi (ed.), *ArcheoFOSS Free, Libre and Open Source Software e Open format nei processi di ricerca archeologica. Atti del VII Workshop (Roma 2012)*, «Archeologia e Calcolatori», Suppl. 4 (http://www.archcalc.cnr.it/indice/Suppl_4/14_Mandolesi_Cocca.pdf).

Marwick B. 2017, *Open science in archaeology*, «The SAA Archaeological Record», 17, 4, 8-14 (https://doi.org/10.17605/OSF.IO/3D6XX).

Moody B., Dye T., May K., Wright H., Buck C. 2021, *Digital chronological data reuse in archaeology: Three case studies with varying purposes and perspectives*, «Journal of Archaeological Science: Reports», 40, 103188 (https://doi.org/10.1016/j.jasrep.2021.103188).

Walsh P., Pollock R., Björgvinsson T., Bennett S., Kariv A., Fowler D. 2018, *Fiscal Data Package* (https://specs.frictionlessdata.io/fiscal-data-package/#language).

ABSTRACT

This paper presents an update to an earlier proposal for a standardized open format for archaeological stratigraphy data, the Harris Matrix Data Package, and the accompanying software tool implementation. The update is two-fold: firstly, it aims at a clear separation between data format and the software tool, particularly by defining the data format in more detail and independently from the software used to create or analyze it; secondly, it introduces a new software feature that allows the creation of a new 'data package' from scratch. A third issue that was identified is the lack of tools for converting existing data to and from the Harris Matrix Data Package, but this issue is not dealt with in this paper.