

KELLIA White Paper on Metadata Standards for Digital Coptic


By Ulrich Schmid, So Miyagawa, Caroline T. Schroeder

From the Koptische/Coptic Electronic Language and Literature International Alliance (KELLIA) Project (<https://kellia.uni-goettingen.de>)

Product of a joint Grant from the National Endowment for the Humanities (HG-229371) and Deutsche Forschungsgemeinschaft (BE 4172/1-1)

Project Directors: Caroline T. Schroeder, University of the Pacific (American PI), Amir Zeldes, Georgetown University (co-PI), Heike Behlmer, Georg-August University, Göttingen; Göttingen Academy of Sciences and Humanities (German PI)

Institutional Grantees: University of the Pacific (NEH), Georgetown University (NEH), Georg-August University (DFG)

The KELLIA White Paper on Metadata Standards for Digital Coptic by [Ulrich Schmid](#), [So Miyagawa](#), [Caroline T. Schroeder](#) is licensed under a [Creative Commons Attribution-NonCommercial 4.0 International License](#). 

27 March 2019

Metadata Standards for Digital Coptic

Members of the Advisory Boards for the Coptic Scriptorium project and the Digital Edition of the Coptic Old Testament project as well as project participants have reviewed this document prior to publication. This White Paper is also attached to the main KELLIA project White Paper (<https://kellia.uni-goettingen.de/downloads/KELLIA-white-paper.pdf>) as Appendix 4.

1. Introduction

Metadata records basic descriptive and administrative metadata like license statements and call numbers from holding institutions, as well as much more detailed descriptions, e.g., of the materiality of information carriers or more abstract aspects of the encoded source and its cultural implications are possible.

German partners completed a survey of Metadata Standards and formats used in the field of Coptic studies and neighboring disciplines. The survey results are in a [database published as KELLIA E-ditions](#).¹ This appendix contains a summary of findings and recommendations based on the survey and work in KELLIA. The survey was conducted *prior* to the establishment of the PATHs project in Rome.² PATHs will be providing unique identifiers to Coptic text-bearing objects; we encourage projects to follow PATHs ongoing work.

2. Encoding

Most projects use TEI-P5 or a specialised subset of TEI as EpiDoc to ensure interoperability in theory. TEI XML (and the EpiDoc subset) includes the msdescription-module with elements for describing manuscripts.³

¹ <http://kellia.uni-goettingen.de/editions/>

² <http://paths.uniroma1.it/>

³ <http://www.tei-c.org/release/doc/tei-p5-doc/en/html/MS.html>

Since every project has its own perspective, in reality many tags are interpreted in different ways due to the very general TEI-Definitions. Projects often develop their own metadata categories and map a set of metadata on to TEI-P5. The Coptic Old Testament project's metadata model is [available online as an example](#).⁴

3. Authority files

Recommended is also the use of authority files or similar data to link identical entities to a general information resource. For Coptic Studies, currently the best solution is to link to Trismegistos People⁵ for a historic person or authority data for proper names. (Note: the Trismegistos People database *is not* a complete or precise prosopography.)

Pelagios⁶ and Pleiades⁷ are the best authorities for linking ancient geographical entities. To find names to specify a geographic location one may use GeoNames⁸ or Getty.

PATHs is developing identifiers for place names in Egypt; projects should contact PATHs regarding their metadata and identifiers for provenance.

A general tendency is to link personal, geographic or other kinds of entities to Wikidata⁹ to use structured linked open data-sets (also known as “Wikification”). One benefit is that one can generate data-sets that can be used by anyone.

4. Standards and Controlled Vocabularies

There are no cross-project controlled vocabularies for metadata, and existing controlled vocabularies (Getty, Dublin Core, EAGLE) do not suffice for Coptic Studies. We recommend that each project publish their own controlled vocabularies to ensure data integrity and consistency within the project; new projects should survey existing projects so they do not need to “reinvent the wheel.” For more on controlled vocabularies, see

⁴ <https://docs.google.com/spreadsheets/d/189qBxOylUyo0rgUSP20kdkE5SBaAAHg7tAerIL1D5Yg>

⁵ www.trismegistos.org

⁶ <http://commons.pelagios.org/>

⁷ <https://pleiades.stoa.org/>

⁸ <http://www.geonames.org/>

⁹ www.wikidata.org

[KELLIA White Paper on Linked Data Standards and Practices](#), also published as Appendix 5 of the main KELLIA project White Paper.¹⁰

The PATHs project in Rome will be publishing stable identifiers for every Coptic manuscript and for place names in Egypt; it will also Clavis Coptica entries for each known Coptic text work.¹¹ ***KELLIA partners encourage all Coptic projects to include these identifiers in their metadata.***

The Coptic Old Testament project explored mapping VMR-Data to METS/MODS to provide an internationally approved metadata exchange format. This task could not be accomplished:

- Mapping the complex VMR structure to the even more complex METS/MODS was time consuming; we aborted the trial during the conception-phase after realising that we do not have the time to dig deep into METS/MODS to achieve a proper result.
- Due to the fragmentation of Coptic manuscripts, one “document” would be split in the VMR into different items with different rightholders and holding institutions. Mapping multiple items to a single METS/MODS representation was difficult, since METS/MODS is designed to present a single dataset for a single legal Resource.

5. Data-Access

To make data accessible for future research the metadata should be saved and provided in a machine-readable form and under special licence agreements that allow re-usage.

Many Coptological projects are accessible via self published or institutionally hosted websites.¹² But it depends on the project holders themselves if and how they publish their data on this platform. On some websites one can easily download the required information because free access to the data is provided. Just to name a few examples,

¹⁰ <https://kellia.uni-goettingen.de/downloads/KELLIA-linked-data-white-paper.pdf>,
<https://kellia.uni-goettingen.de/downloads/KELLIA-white-paper.pdf>

¹¹ <http://paths.uniroma1.it/>

¹² This White Paper’s scope is digital editions projects in a narrower sense (see above). As a side note, the blog of Alin Suciuc should nevertheless be mentioned, where resources regarding coptological (mainly philologically centered) publications are provided. (<http://alinsuciuc.com>)

this is the case for the following projects: Coptic Scriptorium¹³, Inscriptions of Israel / Palestine¹⁴, U.S. Epigraphy Project¹⁵, epidat - epigraphische Datenbank¹⁶, Monasterium¹⁷, Epigraphische Datenbank Heidelberg¹⁸, Inscriptiones Graecae¹⁹, digilibLT - Biblioteca digitale di testi latini tardoantichi²⁰, Bibliotheca Palatina digital²¹, Papyri.info²² and Papyrus und Ostraka Projekt²³. Most of them provide the data via XML-files. The Deutsches Textarchiv²⁴ and Germania Sacra. Die Kirche des Alten Reiches und ihre Institutionen²⁵ even offer more data formats like TCF, Turtle or json-ld). Sometimes a free registration is required to download the information (Papsturkunden des frühen und hohen Mittelalters²⁶) and on a few occasions one can only register by paying a fee (Corpus dei Manoscritti Copti Letterari (CMCL)²⁷).

Several homepages provide a section in which a reference to similar or related projects such as Epigraphische Datenbank Heidelberg²⁸, U.S. Epigraphy Project²⁹, Inscriptions

¹³ <http://copticcriptorium.org/>

¹⁴ <http://cds.library.brown.edu/projects/Inscriptions/index.shtml>

¹⁵ <http://usepigraphy.brown.edu/projects/usep/collections/>

¹⁶ <http://www.steinheim-institut.de/cgi-bin/epidat>

¹⁷ http://monasterium.net/mom/home?_lang=deu

¹⁸ <http://edh-www.adw.uni-heidelberg.de/home?lang=de>

¹⁹ <http://telota.bbaw.de/ig/>

²⁰ <http://digiliblt.lett.unipmn.it/index.php>

²¹ <http://digi.ub.uni-heidelberg.de/de/bpd/index.html>

²² <http://papyri.info/>

²³ <https://papyri.uni-leipzig.de/content/start.xml?XSL.lastPage.SESSION=/content/start.xml>

²⁴ <http://www.deutschestextarchiv.de/>

²⁵ <https://adw-goe.de/fr/forschung/forschungsprojekte-akademienprogramm/germania-sacra/>

²⁶ <http://www.papsturkunden.de/EditMOM/home.do>

²⁷ <http://www.cmcl.it/>

²⁸ <http://edh-www.adw.uni-heidelberg.de/links>

²⁹ <http://usepigraphy.brown.edu/projects/usep/links/>

of Israel / Palestine³⁰, Papyri.info³¹, and Germania Sacra: Die Kirche des Alten Reiches und ihre Institutionen³² can be found. In doing so, it is easier for the user to get to know and browse already existing and theme related undertakings. But this is a far cry from an extensive catalogue in which projects using digital methods are listed as it is just rudimentary linking from project-website to project-website. A central coptological platform aggregating discipline-related digital editions or similar data does not exist at the moment.

Some projects provide access to content data via defined APIs like OAI-PMH. (See Das Altägyptische Totenbuch. Ein Digitales Textzeugenarchiv³³ OAI-PMH interface.)

6. Recommendations

Data created by computer aided Coptological research should be provided as machine-readable and thus digital data. In this manner encoded data including metadata should be archived by institutionally or disciplinary bound repositories which allow long-term digital preservation. This comprises not just storage space or its support but also persistent identification of digital resources via persistent identifiers (PIDs) like DOI³⁴ or other kinds of Uniform Resource Identifiers (URI).³⁵ As an institutional Repository, for example, TextGrid-Repository³⁶ can be mentioned which created a consistent long-term preservation policy and infrastructure: Each digital resource [there] is identified by a PID and accessible via an URL.³⁷ Furthermore the Repository and its data is fulltext searchable and thus provides direct access to all digital data produced. Whereas recommending a specific Repository is not in the scope of this whitepaper's

³⁰ <http://cds.library.brown.edu/projects/Inscriptions/related.shtml>

³¹ <http://papyri.info/docs/resources>

³² <https://adw-goe.de/fr/forschung/forschungsprojekte-akademienprogramm/germania-sacra/links/>

³³ <http://totenbuch.awk.nrw.de/>

³⁴ <https://www.doi.org/>

³⁵ Persistent identifiers are not just to identify resources but also to cite content from digital resources.

³⁶ <https://textgridrep.org/>

³⁷ See for example the resource textgrid:123rw that may be accessed via https://textgridrep.org/browse/-/browse/123rw_0

responsibility, repositories that are qualified by a **DINI-Certificate**³⁸ or a similar certification mark may be more desirable.

Machine-readable data created during the process of digital editing should be explicitly bound to a license which allows free access by means of scientific reuse. Therefore, the Creative Commons Licenses CC-BY³⁹ and CC-BY-SA⁴⁰ are recommended.

Data access does not just concern physical access to analog or digital resources but also legal aspects with regard to creation and usage of resources. In that manner in the metadata, legal technicalities should be clarified for both the source and the digital encoded data regarding authorship, data privacy and personality rights if applicable.

For linking and access, well-resourced projects should consider providing access to metadata via APIs (REST, OAI-PMH). Smaller projects can provide metadata downloadable as csv files for further manipulation and research.

One desiderata is a central Coptological platform that catalogues projects with descriptive and administrative metadata that can be searched (possibly drawn from linked data API's from the projects). Desirable attributes include the following features:

- in English and including projects world wide
- Searchable metadata and documentation:
 - the link to the according project website (maintained to avoid dead links)
 - a brief outline of the project
 - the responsible persons and institutions
 - metadata with regard to geographically, chronologically, thematically, institutionally, data-access (free, registration, fee-based), status (ongoing, completed, discontinued) information
- a defined scope, eg. “digital editions”, “linguistically encoded corpora” etc.

³⁸ <http://www.dini.de/dini-zertifikat/>

³⁹ <https://creativecommons.org/licenses/by/3.0/>

⁴⁰ <https://creativecommons.org/licenses/by-sa/3.0/>

- interactive: the user can contribute to the list of digital projects or correct an entry
- list can be downloaded in various forms
- license for use of data is stated
- connection with other networks / platforms