KELLIA White Paper on Linked Data Standards and Practices for Digital Coptic

By Elizabeth Platte, Caroline T. Schroeder, Ulrich Schmid

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)

27 March 2019

Linked Data Standards and Practices 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 5.

1. Introduction

As part of the overall KELLIA collaboration goal of establishing data standards for related and future projects, KELLIA set out to create standards that ensure collaboration, data-exchange, and compatibility in linked-data initiatives across digital Coptic projects. To this end, we have met with members of other projects working on linked data initiatives in the ancient world, created products that are a result of data exchange and link data across project in the KELLIA collaboration, and considered current practices in linked data when establishing standards, especially for metadata. KELLIA partners and other digital projects working in Coptic need robust linked data infrastructure that allows linking between KELLIA projects and to other projects, even when standards between projects vary.

Several projects currently present opportunities to link geographical data from the ancient world. In particular, US KELLIA partners have begun to work with Pleiades (https://pleiades.stoa.org/), a digital gazetteer of the ancient world. In doing so, they have also worked with Pelagios (http://commons.pelagios.org/), which provides infrastructure for linking geographic data. Pelagios works largely with projects focused on the ancient and medieval periods. Both US and German KELLIA projects have also used Trismegistos (http://www.trismegistos.org/) identifiers in metadata. Trismegistos is a group of metadata databases of information about texts, collections, authors, people, and places. It originally began with information from papyrological and epigraphic information from Egypt, but has since expanded to information from the ancient Mediterranean world more generally.

Finally, KELLIA partners also look forward to working with PAThs (http://paths.uniroma1.it/), a project focused on putting Coptic literary texts into their geographic context. PAThs plans to produce an “Archaeological Atlas of Coptic Literature,” and, in doing so, collect and classify information about Coptic literature, collections, manuscripts, colophons, authors, copyists, donors, and institutions. Unlike Trismegistos, which does not contain a comprehensive catalogue of Coptic manuscripts and manuscript fragments, the PAThs project plans to provide permanent identifiers for Coptic text-bearing objects. Ideally, future work will entail collaborating with Trismegistos to update their catalogues. Such work will be very valuable for KELLIA members, and controlled vocabularies and metadata standards in Coptic digital humanities will certainly be influenced by the work of the PAThs project. Because PAThs is a relatively new project, KELLIA partners are awaiting PAThs outcomes to take advantage of linked data opportunities more fully. Sharing and linking data makes the work of the individual projects and initiatives that are part of KELLIA more valuable, as other projects use and expand it. This has already been demonstrated by outcomes of sharing data noted below. Linking data to projects both within and outside of KELLIA also makes our individual work more discoverable.

2. KELLIA Linked Data Products

Online Coptic Dictionary: Entries in the online dictionary (http://coptic-dictionary.org) are linked to Coptic SCRIPTORIUM digital textual data (described in more detail in the main KELLIA White Paper Grant Products section). The same linking opportunities are available to KELLIA partners and other Coptic-language projects, who can freely annotate their digital texts with links to online dictionary lemma searches.

Coptic Treebank: The automatic syntactic annotation lays the groundwork for linking data about entities in Coptic texts, as described below. The treebank standards themselves are described in more detail in the main KELLIA White Paper Grant Products section.

Entity recognition: Proof-of-concept machine-processed entity-recognition (described in more detail in the main KELLIA White Paper Grant Products section) will allow for entity disambiguation and linking opportunities for Coptic literature to named entities projects such as Pelagios, Pleiades, Trismegistos, PATHs.

VMR-Coptic SCRIPTORIUM Converter: This converter (described in more detail in the main KELLIA White Paper Grant Products section) transforms digital text data produced by the Coptic Old Testament project’s Virtual Manuscript Room to the EpiDoc TEI XML format used by Coptic SCRIPTORIUM and pushes this XML text to Coptic SCRIPTORIUM’s natural language processing pipeline. This converter facilitates the integration of digital text corpora developing in both projects.

3. Recommended standards

The KELLIA project recommends the following linked data standards, described in more detail below.

- Projects should adopt and publicly document controlled vocabularies for use in metadata, thus ensuring data integrity and the possibility of linking via API (section 3a)
- Projects should provide automated metadata validation whenever possible (section 3a)
- Projects should include metadata fields for collaborators' unique identifiers (section 3a)
- Projects should assign persistent identifiers to the digital objects they produce (section 3b)
- Projects should include clear licensing information to facilitate data exchange (section 3b)

3a Standards for controlled vocabulary

In order to facilitate retrieving, linking, and exchanging data, KELLIA partners have agreed upon recommended metadata and data structures, as outlined in the metadata standards appendix of this grant report. We recommend all digital Coptic projects establish internal controlled vocabularies for metadata. Such vocabularies provide internal standards for populating metadata fields to ensure data integrity. Additionally, in the absence of a robust hub for linked data (akin to Pelagios or Trismegistos), controlled vocabularies lay the groundwork for linking via API's or queries rendered in http URI syntax. (E.g., in Coptic Scriptorium's environment, the query http://data.copticscriptorium.org/filter/author=Shenoute retrieves all documents where the "author" metadatum is "Shenoute", and http://data.copticscriptorium.org/filter/collection=Borgia%20Collection retrieves all documents where the "collection metadatum is "Borgia Collection"). Controlled vocabularies accessed through URI based queries can enable a project to achieve 4-star level linked data. Well-documented internal standards can also help to facilitate the exchange of data, as conversions can be applied to communicate across different project standards. Standards thus lay the groundwork for a more robust linking environment for Coptic data in the future (such as Pelagios is for ancient geography).

To date, standard controlled vocabularies (such as Dublin Core, Getty, or the Europeana EAGLE Vocabularies[1]) do not provide the coverage needed for Coptic literary materials, and there is no comprehensive standard for controlled vocabulary for fields such as places, names, document identification, etc. in digital Coptic. Despite this lack of standard, we recommend projects develop internal standards, and the metadata guidelines in this White Paper provide models. We also recommend an automated process for metadata validation to ensure adherence to internal standards whenever possible. Our current annotation editor, the open source online XML editor GitDox, provides automatic checking of metadata field names and values, ensuring that required metadata appears and matches regular expression patterns specifying valid values across our documents. Likewise, whenever possible, projects should include unique identifiers assigned to metadata fields by other projects. KELLIA partners should therefore include a field for CoptOT/Coptic SCRIPTORIUM identifiers in metadata.

The Coptic Old Testament project has created and shared a list of standard names for institutions that have Coptic manuscripts in order to standardize their metadata. Their list can be found here: (https://docs.google.com/spreadsheets/d/1jDeRLhM9tlG9pzkUWbyoi7CD2_L4yLdDM1yJ18ToT9A/edit#gid=0). It includes references to the respective entries in the Trismegistos “Collections” table http://www.trismegistos.org/coll/index.php Such standards will make the discovery and exchange of data across projects easier.

3b Standards for unique identifiers

As mentioned in the recommendations for data curation, digital objects created by each projects should be assigned persistent identifiers, such as Digital Object Identifiers or other Unique Resource Identifiers that are accessible via a URL (1.3.3). Section on KELLIA current practices. Coptic SCRIPTORIUM assigns a unique document title to each document as well as CTS (Canonical Text Service) URNs (http://www.homermultitext.org/hmt-doc/cite/cts-urn-overview.html) for the digital data in their corpora. The URN schema includes two namespaces created specifically for Coptic materials: copticLit for literary materials and copticDoc for documentary materials. Coptic SCRIPTORIUM provides a service (data.copticscriptorium.org) to resolve these URNs, and they can also be used in the ANNIS search and visualization interface.

Finally, as mentioned in the recommendations for data curation standards, all projects should have clear licensing attached to all products in order to facilitate data sharing

[1] https://www.eagle-network.eu/resources/vocabularies/

and linking (1.3.4). Ideally, this information should both be included in the metadata associated with each digital object and be evident or easily findable through any web-based interface used to display the object.

4 Future plans in Coptic linked data

KELLIA partners look forward to continuing to share data and standards, as well as collaborating with other Coptic and ancient studies projects. We have already begun to work with Pelagios and Pleiades to link our geographic data and make it discoverable to a wider scholarly and popular audience. We are especially excited about working with the PAThs project in the future, as their proposed interlinked databases will provide an opportunity for other Coptic projects to review data and metadata standards in relation to their work.