RESEARCH PAPER

Applying the Canonical Text Services Model to the Coptic SCRIPTORIUM

Bridget Almas[1] and Caroline T. Schroeder[2]

[1] Tufts University, US

[2] University of the Pacific, US

Corresponding author: Bridget Almas (balmas@gmail.com)

Coptic SCRIPTORIUM is a platform for interdisciplinary and computational research in Coptic texts and linguistics. The purpose of this project was to research and implement a system of stable identification for the texts and linguistic data objects in Coptic SCRIPTORIUM to facilitate their citation and reuse. We began the project with a preferred solution, the Canonical Text Services URN model, which we validated for suitability for the corpus and compared it to other approaches, including HTTP URLs and Handles. The process of applying the CTS model to Coptic SCRIPTORIUM required an in-depth analysis that took into account the domain-specific scholarly research and citation practices, the structure of the textual data, and the data management workflow.

Overview

Coptic SCRIPTORIUM (SCRIPTORIUM) is a platform for interdisciplinary and computational research in Coptic linguistics, literature, and history. The goal of this research project was to investigate, decide upon and implement a system of stable identification for the texts and research objects in SCRIPTORIUM in order facilitate their citation and reuse.

SCRIPTORIUM began when two scholars from different disciplines (Amir Zeldes in Linguistics and Caroline T. Schroeder in Religious Studies) identified a shared need for digitized corpora of Coptic texts. The Coptic language was the last phase of the ancient Egyptian language family, in use from late antique to Byzantine periods of Egyptian history. Due to Egypt's dry climate, Coptic sources important to a number of academic disciplines have survived. Linguists conducting computational and statistical research into language as well as historians, religious studies scholars, and philologists seeking to query and investigate topics and terminology across large sets of primary sources would find a digital corpus annotated consistently according to recognized standards useful. Through SCRIPTORIUM, Zeldes and Schroeder began digitizing primarily Coptic literary texts (such as formal letters, sermons, gnomic sayings, and monastic rules) composed primarily by monks of the fourth through sixth centuries and copied by later scribes through approximately the twelfth century. In addition, SCRIPTORIUM has developed tools and standards for annotating Coptic literature for linguistic, paleographical, and historical information, as well as an online digital environment for reading and querying the texts, translations, and annotations. Identifiers for texts, annotations, and other features are necessary.

Our basic requirements for the SCRIPTORIUM data identifiers were that they be stable, location and technology independent, and globally unique. We wanted an identifier scheme that would outlast any particular delivery mechanism, such as the web, but also required that the identifiers be resolvable in the context of the systems that do exist today. In particular, the SCRIPTORIUM data must be able to be cited as linked data on the web according to the principles of "5 Star Linked Data."(Berners-Lee, 2006)

Our analysis took into account the specifics of the SCRIPTORIUM data resources and its current data production and management practices as well as the domain-specific scholarly practices of citation of the texts in the SCRIPTORIUM corpus.

Scholarly Citation Practices

Scholarly citation practices in Coptic studies vary, due in large part to the dispersed geography of the corpora and the diversity of scholarly fields using Coptic sources. The history of collecting, editing, and publishing Coptic literary texts in print has affected scholarly citation practices.

Manuscripts containing Coptic literary texts (whether parchment or papyri, single pages or codices) now reside in multiple repositories primarily in Europe and North America. One codex or papyrus may have been broken into pieces at some point during its history and thus survive in a dismembered state in more than one repository. In other words, multiple libraries or museums may each possess a different part of the same codex; one letter or sermon or biblical book may be split geographically. Consequently, publication of these manuscripts has been spotty. At times (such as in the case of Marius Chaîne's edition of the Coptic *Apophthegmata Patrum,* or *Sayings of the Desert Fathers*, hereafter AP), a print edition has included all known manuscripts of a work or a collection of works (even when those manuscripts are in multiple repositories), in the order in which the manuscript pages originally were composed, and the editor has a imposed a clear numbering scheme to divide the work(s) into easily citable segments. In other cases, such as the works of the monk Shenoute, editors have published manuscript fragments known to them due to their familiarity with certain libraries and collections, but these fragments are neither in order nor comprehensive; fragments pertaining to the same conceptual work may not be published one after the other, texts on some fragments were erroneously attributed to Shenoute, and many manuscripts remain unpublished.[1]

As a result, some Coptic literary texts have existing functional, canonical referencing systems for print editions (such as Chaîne's *AP*), and others (such as Shenoute's corpus) do not. For those corpora without comprehensive print editions, older citation practices have included referencing the page number and book for published texts and the current repository shelf mark for unpublished texts. While this practice enables later scholars to find the original fragment quoted, it does not help readers locate the conceptual work cited and the other fragments of that same conceptual work. More recently with Shenoute's corpus, scholars have begun to cite documents using the incipit of the original conceptual work (e.g. Shenoute's sermon *I See Your Eagerness*) following a list of incipits published in Stephen Emmel's codicological reconstruction of the corpus (Emmel, 2004). Scholars have also begun citing some combination of the original manuscript codex and pagination and current repository shelf mark as well as print edition and page numbers. For many Coptic literary texts (including Shenoute), canonical chapter or verse divisions in works do not exist.

Citation practices vary across scholarly fields, as well. Papyrologists cite Coptic papyri, ostraca, and other text-bearing objects using a community-edited Checklist of Editions (n.d.), which lists abbreviations for documents classified based on a variety of criteria, including the type of text-bearing object, the original location of the text-bearing objects (e.g., P.KellisCopt for Coptic papyri from Kellis), or the repository holding the text-bearing objects at the time they were published (e.g. P.Köln for papyri located in Cologne and O.Lips. Copt. for Coptic ostraca in Leipzig). The abbreviations in the Checklist are standards, used internationally across the field and used as metadata or sometimes even in the identifiers for digital editions published at papyri.info.[2] These abbreviations are functionally canonical citations for the field, though not without complications; for example, some documents may have been published more than once under multiple sigla, or documents in one collection may be fragments related to fragmentary documents in another (as we saw with the Coptic literary texts).

Despite the diversity of approaches to citation, most involve some sort of semantically meaningful identifiers for the texts and passages being cited. We felt therefore that retaining some element of semantic meaning for our machine-actionable identifiers, while not a technical requirement, was an important cultural one. Scholars in this field do not typically cite databases employing non-semantic identifiers; even papyri.info's identifiers have semantic meaning, as we have shown. Semantic identifiers allow others to find specific documents cited and related documents with greater ease.

We asked a sampling of our expected audience of researchers about their anticipated citation needs and practices when using the digital corpus. (The sample consisted of six scholars from two different disciplines and different career stages.) From this we identified the following research activities and citation targets as our core user stories (**Table 1**). The user stories cover of the entire SCRIPTORIUM resource set, but only a subset apply to our initial scope.[3]

---

[1] For more on the ways the manuscript history affects digitization, see Schroeder and Zeldes, 2016.

[2] E.g., http://papyri.info/ddbdp/o.petr.mus;;554

[3] User stories were provided by researchers in natural language, and were reformatted according to the above standard syntax for purposes of clarity and consistency for this paper.

ID	User Story	In Scope?
1	I want to use the HTML normalized visualizations to read a text and be able to capture the identifier of the specific text and visualization I used so that I can cite it in a publication.	Yes
2	I want to use the HTML normalized visualizations or analytic visualizations to read and cite a text and also refer be able to capture the location of the diplomatic visualization so that I can check something in the text.	Yes
3	I want to be able to cite a word search conducted in normalized layer in a word study of a particular corpus or author's work.	No
4	I want to search for all forms of a dictionary headword (example, the word for "destroy") in order to study how it gets used in different forms in a particular text, and then be able to cite all documents in a particular group with this word in it, including the original and normalized layers.	Yes
5	I want to search for a particular part of speech tag or tags and download and cite all phrases (i.e. several tokens surrounding the tagged word) containing that part of speech.	Partially
6	I want to search for loan words using the language annotation and download and cite the data in original and normalized form, including the language and part of speech annotations.	Partially
7	I want to be able to search for N-grams with a specific tag in order to analyze style and then cite the corpora in which they are found.	Yes
8	I want to be able to cite biblical verses quoted by a specific author or author group so that I can cross-reference them in my other corpora.	Yes
9	I want to be able to cite search results for certain markers for scriptural citation so that I can isolate and compare uses of certain rhetorical phrasing in a specific text.	Partially
10	I would like to be able to use the static HTML visualizations for a handout for a conference paper.	Yes
11	I want to cite specific passages in the texts.	Partially
12	I want to cite specific passages with an equivalent specificity to page and /line numbers.	Partially

Table 1: User Stories.

Identifiable Resources

The SCRIPTORIUM corpus includes a wide range of resources, both abstract and concrete, that are uniquely identifiable and citable, whether as a whole or some selected fragment, part or range (**Table 2**).

Ultimately each and every addressable resource in the SCRIPTORIUM corpus, whether a source text, an annotation, or other related resource, should be assigned a stable, resolvable identifier. Although we initially hoped to include the entire data set in the scope of this project, the complexity and diversity of the data required us to limit the scope to include only the conceptual works and their digital expressions, postponing the rest for a future phase of work. The uses cases for dealing with finer-grained parts of texts, such as word tokens and bound groups, call for collection models and services such as those currently being defined by the Research Data Collections Working Group of the Research Data Alliance (RDA)[4]. Coptic SCRIPTORIUM is use case provider and potential adopter of this currently ongoing effort.

Delving further into the refined scope, we found that the category of "digital expressions" needed to be expanded to include the distinct representations (or visualizations) of each expression (**Table 3**). The digital expressions are text transcriptions, annotations, visualizations of texts and/or annotations, and data files in different formats that store or express the text and annotations.

Analysis
CTS URN Specification

We began the project with an assumption that we would use the Canonical Text Services (CTS) model, developed by the Homer Multitext Project at Harvard's Center for Hellenic Studies, for our textual identifiers. The CTS specification defines a URN based identifier structure for identifying texts and canonically

---

[4] https://rd-alliance.org/groups/pid-collections-wg.html

Resource Type	Description
Conceptual works	Letters, treatises, sermons, contracts, monastic rules and other texts, as distinguished from the individual instantiations, editions, or versions of these "conceptual works" that survive in specific manuscript witnesses and publications.
Original physical manuscripts representing instantiations of the works	Fragmented papyri and parchment codices, housed in different museum, library, and private repositories.
Digital versions of the physical manuscripts	Digital facsimiles such as photographs and/or existing digital transcriptions of texts.
Logical groupings of works	Author groups (e.g., texts by monastic authors Shenoute or Besa) or groupings in which anonymous or multi-authored works have circulated or been grouped historically (e.g., the "New Testament" or collections of sayings known as the *Apophthegmata Patrum*).
Physical collections housing the manuscripts	Museum, library, and private repositories with unique (and sometimes changing) cataloguing systems.
Physical codices, each containing one or more manuscripts	Books (fragmentary or whole) that may contain multiple works; the works may have been bound into a codex in the original, ancient repository (such as a monastery), or works (fragmentary or whole) may have been bound into a codex at a modern repository, having been collected from different original sources but now all residing in a modern collection.
Authors of works	Original author, named or anonymous.
Scribes or hands	Persons who transcribed the text on the existing witness; may or may not be the author (in literary texts rarely if ever the author).
Paleogeographic symbols	Notations including punctuation, supralinear strokes.
Digital images of manuscripts	Digital photographs of manuscript pages and papyri, of varying quality and accessibility depending on the repository; majority of documents not photographed or photographs unavailable to the general public.
Linguistic Annotations[5]	Annotations made by modern editors for linguistic properties, such as part of speech (nouns, verbs, articles, etc.) or syntax.
Morphemes[6]	Linguistic units smaller than words that are linguistically or lexically meaningful and therefore annotated.
Word tokens	Coptic words.
Bound groups	Coptic is an agglutinative language with words and morphemes joining together in groupings (e.g., subject pronoun + verb + direct object pronoun).
Editors and Annotators of the Coptic Scriptorium project	Names of individuals who have transcribed and annotated documents.

Table 2: Coptic SCRIPTORIUM Resources.

cited passages of texts.[7] There is also a companion CTS Application Programming Interface (API) protocol for a service to retrieve fragments of texts by canonical reference, as expressed by their CTS URNs[8]. The CTS model is both human and machine-readable, and enables identification of digital editions and visualizations of previously published as well as unpublished documents.

The CTS URN model partially overlaps with the Functional Requirements for Bibliographic Records (FRBR) model. CTS derives from the conceptual entity-relationship model of FRBR, to describe the editions and

---

[5] For principles and best practices for part of speech tagging in Coptic, see Zeldes and Schroeder, 2015; for regularly updated guidelines see Zeldes and Schroeder, "SCRIPTORIUM Part-of-Speech Tagsets for Sahidic Coptic"; For lemmatization guidelines, see Zeldes, "Coptic SCRIPTORIUM – Lemmatization Guidelines."

[6] On the implications of Coptic grammar and word segmentation for digitization (especially the concepts of morphemes, words, and bound groups), see Schroeder and Zeldes, "Raiders of the Lost Corpus."

[7] http://cite-architecture.github.io/ctsurn_spec/

[8] We limited the scope of this effort to the application of the CTS URN specification and did not deploy an instance of the CTS API for serving textual passages. The latter may be considered for a future phase of work.

Visualization/Representation	Description
Normalized HTML	An HTML visualization of text annotations in which the Coptic text has been normalized for spelling, punctuation, word segmentation for ease of reading.
Analytic HTML	An HTML visualization that expresses multiple annotations, particularly by aligning a normalized Coptic annotation, an English translation annotation, and the part of speech annotation.
Diplomatic HTML	An HTML visualization of a transcription that most closely resembles a manuscript transcription. Coptic text annotations include original spellings, punctuation, orthography; visualization also may also express annotations for the manuscript's line, column, and page breaks, ink color, gaps or lacunae, and other paleogeographic information (Krause and Zeldes, 2013).
TEI XML	Coptic text and annotations encoded in XML according to the EpiDoc subset of the Text Encoding Initiative standards[9]. (Does not contain the full set of linguistic and syntactic annotations.)
PAULA XML	Coptic text and annotations encoded in standoff markup XML according to PAULA standards[10]. Contains full set of annotations.
relANNIS	Relational database files including all text and annotations for search and visualizations in the ANNIS[11] database infrastructure.
ANNIS Visualization	Live search and visualizations of the texts and annotations in the ANNIS infrastructure installed on a web server.

Table 3: Citable Visualizations.

translations as expressions of a work and online instances of the text as items (it skips the FRBR concept of manifestation). But it also extends FRBR to support and ensure, within a single model, a rigorous method of identifying texts as citable nodes in an "ordered hierarchy of content objects" (OHCO) as first proposed by Renear, et. al. (1992). In the CTS model, the texts themselves are positioned within a hierarchy, as per FRBR, such as belonging to a collection of works by a specific author or group of authors, but also serve as containers for citable hierarchical nodes (i.e. passages). The CTS specification declares that stable identifiers must support the following requirements of a theoretical model: (1) identifiers must be abstract, machine-actionable and technology independent; (2) citable texts consist of a set of citable nodes; (3) there are four properties of a citable node: (a) it belongs to specific version of a work in a *FRBR-like* hierarchy; (b) it belongs to a citation hierarchy of 1 or more levels; (c) is ordered; (d) it may have mixed content; (3) identifiers must be immutable (but may be versioned). (Smith and Blackwell, 2012).

As discussed in the Scholarly Citation Practices section of this document, semantically meaningful identifiers were an important requirement for this project. CTS URNs function to identify documents and text groups and can also enable identification of the most recent version of those documents or texts groups.

CTS URNs are composed of the following colon-separated parts:

```
urn:cts:NAMESPACE:TEXTGROUP[:WORK[:PASSAGE]]
```

The *NAMESPACE* component defines the Naming Authority for a specific set of URN identifiers. The components which following the Naming Authority are declared as unique within that space.

The *TEXTGROUP* component is an identifier for any group of texts that are conventionally cited together "in the naming authority's tradition". Author is the most common grouping concept, but the specification declares this to be an editorial decision.

The *TEXTGROUP* may be followed by *WORK* component. This component is itself made up of multiple dot-separated parts:

```
work[.version[.exemplar]]
```

---

[9] Elliott, Bodard, Cayless *et al.* (2006–2016)

[10] https://www.sfb632.uni-potsdam.de/en/paula.html

[11] http://corpus-tools.org/annis/

The *work* component identifies a specific notional (or abstract) work within the *TEXTGROUP*. The specification states that this component of the URN corresponds to a *FRBR Work*.

The *work* component may be followed by an identifier for a specific *version* of that work, either a translation or an edition. The specification intends this to correspond to the *FRBR Expression*. The *version* component may be followed by an identifier for a specific *exemplar* of the *version*, which the specification intends to correspond to the *FRBR Item*.

CTS URNs may also include a final component, the *PASSAGE* identifier, that identifies either a citable node, or a range of citable nodes, within the text. This may be included for any CTS URN which defines at least an abstract work, so that it is possible to cite the abstract concept of canonical passage within a work, without referring to a specific expression of that work.

The Perseus Digital Library (PDL) published the following URN to identify the canonically cited, abstract concept of Book 1, Line 1 of Homer's *Iliad:*

```
urn:cts:greekLit:tlg0012.tlg001:1.1
```

As well as this URN to identify of Book 1, Line 1, of Homer's *Iliad* in a specific edition, *perseus-grc1,* published by the PDL:

```
urn:cts:greekLit:tlg0012.tlg001.perseus-grc1:1.1
```

Prior to the SCRIPTORIUM effort, the CTS model had not, so far as we were aware, been applied to a corpus without a long tradition of canonical citation, or to texts which followed a fragmentary and non-XML based digital curation model, and thus its' suitability for the SCRIPTORIUM corpus was not certain. The CTS model was designed specifically to make the scholarly tradition of expressing existing canonical citation schemes machine actionable, but it doesn't provide guidance on principles for defining new schemes or best practices for ensuring interoperability across projects (Kavelsmaki, 2014).

Evaluating Existing Identifier Sources

Orthogonal to the analysis of the applicability of CTS, was the question of whether or not to reuse or reference any aspects of existing identifiers for the texts in the corpus. Doing so might enable us to provide semantic meaning that would make the objects of the identifiers immediately recognizable to the scholarly community even without accompanying metadata.

The manuscript sigla developed by Tito Orlandi and the *C*orpus dei Manoscritti Copti Letterari[12] are six-letter sigla identifying ancient codices (e.g., MONB.YA). The first four letters indicate the original ancient repository (e.g., MONB for White Monastery); the second two provide the unique identifier for a codex from that repository. These sigla allow scholars to provide a single identifier to a codex that may now be divided among different modern repositories, with multiple modern shelf marks.

The Clavis Coptica Index[13] is a list of numeric identifiers for authors and works designed by the CMCL as a Coptic version of the *Clavis Patrum Graecorum*.[14] Although extensive, listing 100 authors and far more anonymous works, it remains incomplete.

Trismegistos (Schirrwagen, 2014)[15] (hereafter TM) assigns unique identifiers to ancient text-bearing objects, including codices and papyri, in multiple languages. Similar in concept to the CMCL manuscript sigla but much broader in linguistic, temporal, and geographic scope, with an online database of over 600,000 records. The coverage of medieval Coptic manuscripts, however, is not as comprehensive as the CMCL sigla.

The Perseus Digital Library, through the Perseus Catalog, defines CTS URNS[16] for works that were originally transmitted in Greek in the perseus:greekLit namespace (Almas, Babeu and Krohn, 2014). The Perseus Catalog was conceived in 2005 as a way to integrate two complementary kinds of resources, bibliographies of authors and editions produced by and for classicists and metadata about Greek and Latin authors in more general library systems. The Perseus Catalog builds on the CTS and CITE architecture and has assigned approximately 14,000 CTS URNs to its catalogued primary source editions and translations.

---

[12] http://www.cmcl.it/

[13] http://www.cmcl.it/~cmcl/chiam_clavis.html

[14] Geerard, et al. (1974–2003).

[15] http://www.trismegistos.org/

[16] http://catalog.perseus.org

Papyri.info[17] aggregates digital editions from other existing databases and contributions via its own online papyrological editor. Its over 150,000 identifiers include invoice identifiers and catalogue record identifiers. Identifiers derive from contributing database identifiers or from sigla in the previously cited Checklist of Editions.

There are also various numbering systems applied by editors of print editions that are now integrated into scholarly citation practices, such as numbers Chaîne assigned to individual gnomic sayings in his edition of the Coptic *Sayings of the Desert Fathers*. (Chaîne 1960)

We chose not to reuse the Clavis Coptica identifiers for the SCRIPTORIUM corpus to avoid introducing ambiguity about the source of the data. This decision was informed by the experience of the PDL. For the PDL CTS URNs, the choice to reuse identifiers from the Thesaurus Linguae Graecae (TLG) and Packard Humanities Institute (PHI) digitized canons[18] in the components of the Perseus URNs led to mixed results. That decision was motivated by the assumption that since these were such well known identifiers, they would provide an easy way for people to identify which author and work were represented by the URN. This turned out to be true to some extent, but also had an unexpected drawback in that people were confused about whether or not the texts represented specific editions from those content providers, as evidenced by numerous queries the PDL received asking if and how the identifiers related or didn't to those of the TLG and PHI.

We ruled out reuse of the Trismegistos (TM) Identifiers because only a portion of the SCRIPTORIUM corpus had TM ids and it would have introduced an external dependency to request issuance of new identifiers. (Due to TM's focus on text-bearing objects from 800 BCE to 800 CE, medieval Coptic manuscripts come into the database only when contributed from other partners. It is desirable to work with TM in the future to produce more TM numbers for Coptic codices.) In addition, not all of the texts in the SCRIPTORIUM corpus were in the realm of epigraphy and papyri. The same decision applied to the Papyri.info identifiers, as only a fraction of the SCRIPTORIUM corpus would be papyri or ostraca.

We were tempted to reuse the CTS URNs defined by the PDL for one of the groups of works in the SCRIPTORIUM corpus, the *Apophthegmata Patrum* (AP). These works occur in multiple languages, including Greek. We eventually ruled this option out, as the scholarly preference on the AP was that the Coptic AP should not be treated as derivative of the Greek. Instead, we used the numbering system used by Marius Chaîne, the scholar who edited and published in print the most well-known collection of Coptic AP.[19]

We also reused the Orlandi sigla identifiers in the SCRIPTORIUM identifiers for Manuscript codices. Scholarly print citation practices widely make use of them, and they allow for the identification of now geographically dispersed pages that initially belonged to the same physical codex. Moreover, for Coptic literature they provide a more comprehensive existing list of identifiers than TM numbers.

Applying CTS
Namespace
We chose the *namespace* "copticLit" for URNs pertaining to Coptic literary works. These include hagiography (saints' lives or martyrology), monastic literature, formal letters, and biblical texts. We chose the *namespace* "copticDoc" for URNS pertaining to documentary sources: wills, contracts, and letters typically found on papyri or pot sherds (ostraca). We followed existing disciplinary conventions that have categorized surviving text objects into "literature" and "documentary sources," even though we recognize that these categories are historical, scholarly constructions. We anticipate there may be fluidity in categorizing some text objects as "documentary" or "literary." Texts that problematize these categories would be a biblical quotation found on an ostracon, or a letter that could be considered either literary or documentary. Despite concerns about reinscribing existing, and perhaps problematic, disciplinary boundaries, we nonetheless anticipated that some form of classification would be helpful to users and chose to adapt existing standards.

Textgroups
For the *textgroups*, grouping by attributed author was a logical approach for some of the texts:

```
urn:cts:copticLit:shenoute – texts attributed to Shenoute
urn:cts:copticLit:besa – the letters of Besa
```

However, not all of the corpora could be easily attributed to a single author.

The *New Testament* contains texts of known authorship, contested authorship, anonymous authorship, and multiple authorship. We considered following the approach of the PDL, grouping this under the TLG identifier for this group of texts:

```
urn:cts:greekLit:tlg0031
```

One argument for reuse of the PDL identifiers would be that a machine or person could make some logical assumptions about a "sameAs" type of relationship between the versions of these works in the PDL and relationships. But inclusion of a different namespace in the corpus added confusion, and to follow the true meaning of the namespace, it would require the PDL assume the role of naming authority for this subset of the SCRIPTORIUM texts. Further, as described above, we had concerns as well about reuse of the canonical TLG identifier based upon the experience of the PDL. These considerations led us to define new identifiers that referenced the New Testament semantically:

```
urn:cts:copticLit:nt
```

This choice does perpetuate a notion of a fixed New Testament canon into our classification system, even though scholarship on religious literature in late antique Egypt demonstrates that understandings of sacred literature were remarkably fluid. Even after Athanasius, Archbishop of Alexandria, distributed a letter in 367 positing as canon the texts currently considered to be "the New Testament", evidence shows that Christians throughout Egypt had diverse reading and liturgical practices.[20] Nonetheless, as with some other situations outlined in this essay, we chose an identification scheme that would be recognizable to users. Our decision was also motivated by an existing digital edition of the "Coptic New Testament," which SCRIPTORIUM included in the corpora.

The AP presented a different challenge. These sayings are attributed to multiple authors, and their origins are deeply contested. Many of the sayings also appear in Greek, but not necessarily in the same order. And not all of the Greek sayings appear in Coptic. We chose to group these together in a single *textgroup*:

```
urn:cts:copticLit:ap
```

Works

When it came to defining the *work* component, some texts offered straightforward choices, and others were more complicated.

The conceptual works of Shenoute and Letters of Besa were fairly unambiguous to identify. E.g. we have:

```
urn:cts:copticLit:shenoute.a22 – Acephalous Work 22
urn:cts:copticLit:shenoute.abraham – Abraham our Father
urn:cts:copticLit:besa.aphthonia – Letter to Apthonia
urn:cts:copticLit:besa.thieving – Letter to Thieving Nuns
```

The books of the New Testament are also canonically identified themselves as individual abstract works:

```
urn:cts:copticLit:nt.mark
```

The AP again was the most complex. The individual sayings could be considered works themselves, or referenced as passages within a larger composite work, with identification, and attribution of authorship treated as annotations on those passages. We chose the former approach, making each saying its own conceptual work:

```
urn:cts:copticLit:ap.5
urn:cts:copticLit:ap.6
```

---

[20] See, for example, chapter 1 of Lewis (2013) on the Nag Hammadi Library, and other apocryphal literature from Egypt published in Ehrman and Pleše (2011).

Editions

The most challenging task in the analysis of the hierarchy was determining what comprised a distinct *edition* (the CTS *version* level) and how we wanted to label these.

One question was whether to represent digitizations of different physical codex fragments as distinct editions of a work, or instead to identify editions which represented an aggregate of the folios from extant physical manifestations. There were both scholarly and logistical considerations to take into account.

On the practical side, we needed to consider the logistics of SCRIPTORIUM's curation and publication workflow. Digital texts in SCRIPTORIUM are organized by contiguous manuscript folios in a physical collection rather than as a single file with transcriptions of multiple separate extant fragments. There are various reasons for this approach, including:

· the publication and copyright status of the texts differs from (physical) page to page
· the online publication process can be more agile than it would be if all fragments had to be assembled in a single file before publication
· ANNIS, the linguistic annotation search environment, performs better with smaller documents (Krause and Zeldes, 2014)

This approach to corpus organization is optimized for curation and publication, but added a layer of complexity to the assignment of edition identifiers for the data. We want to be able to identify the folio fragments individually, but without losing the context of the parent work to which they belonged.

From the perspective of the CTS model, there were two options for handling this. We could assign an edition identifier to an aggregate of the digital representations of the folio fragments, and then use the passage component to identify a specific fragment within the edition. Or we could assign an edition identifier to each fragment, giving us multiple editions of a work where each represented a different fragment of that work.

Taking the former approach would mean that we were asserting the folio structure as our canonical citation scheme for the texts. At first analysis, this was appealing, because for some of these works, such as Shenoute's texts, there is no long tradition of citing anything other than the manuscript folios or the page numbers of published print editions. Thus the passage identifiers would correspond with scholarly practice. But a deeper look at the data and the CTS model revealed that this would counter a benefit we hoped to obtain from using the CTS model. One of the features of the CTS URN syntax is that it allows for identifiers for a passage of a notional work, without reference to a specific edition. (As stated in our earlier example, we can use the urn `urn:cts:greekLit:tlg0012.tlg001:1.1` to identify the abstract concept of Homer's Iliad Book 1, Line 1, and urn `urn:cts:greekLit:tlg0012.tlg001.perseus-grc1:1.1` to identify Homer's Iliad Book 1, Line 1 in the specific *perseus-grc1* edition of that work.) But for the fragmentary texts in the SCRIPTORIUM corpus, the folios could not be considered canonically citable under this model, because Folio A in one codex containing the work is not guaranteed to correspond to Folio A in another codex containing the same work.

A further consideration on this point was the fact that the CTS *TextInventory* metadata structure for expressing the contents of a CTS Repository of texts did not provide a means to allow for single editions of a text to be made up of multiple files. While support for this might be a reasonable enhancement request to the CTS specification, this, combined with the passage citation issues, led us to the decision to assign edition identifiers to each fragment. We could use metadata to express relationships between different fragmentary editions.

Another decision point for assigning edition identifiers was whether to consider the different visualizations of a text (see **Table 2**) to be distinct editions of a text. Because each visualization was produced from the same core data set for a text, we chose to follow linked data best practices here, and treat the different visualizations as different *formats* of a single edition.

The final question affecting the scope of the edition identifiers was how to treat translations. The CTS version component has two instantiations, the edition and the translation. Per the CTS model, an edition must be an expression in the original source language of the work, and expressions in other languages are translations. However, in the SCRIPTORIUM curation workflow, translations are produced as annotations on the source text, rather than as separate digitizations. At the time of the project, there was no plan to extract these annotations into separate expressions of a work, so we did not create CTS URNs to identify translations.

Having decided *what* to identify as editions, we needed to determine what to use for this component of the identifiers. We wanted to use semantically meaningful names, and this is where the reuse of existing identifiers played the most significant role.

For the New Testament, we chose to reference the source of our digitized version, J. Warren Wells' "Sahidica"
New Testament[21]

```
urn:cts:copticLit:nt:mark.sahidica
```

For the Besa Letters and the Shenoute and AP texts, we reused the Orlandi Sigla for the manuscripts from
which each edition was derived. However, as per the above-referenced discussion, each edition only con-
tained a set of contiguous fragments from a manuscript, and we would have multiple editions derived from
each manuscript, we also needed to reference the folio pages in these identifiers:

```
urn:cts:copticLit:shenoute.a22.monbya_421_428
```

Exemplars

The CTS Specification does not give explicit guidance on how to use the *exemplar* component of the CTS
URN identifier in a digital environment. It is equated to the *FRBR Item*. One requirement for our identifiers
was that they have the capability to identify distinct versions. In theory, particularly if the identifiers are used
as persistent identifiers, each change to a digital CTS edition should require a new identifier. For a corpus
which is growing and which is mid-curation, minting new URNs for each text would at a minimum require
developing automated tools to manage. Although that was not feasible within the scope of our project, we
did foresee this eventuality and decided to reserve use of the exemplar component of the URN to express
distinct versions of each edition.

When versioning is implemented, the CTS resolver for SCRIPTORIUM will always redirect requests for a
specific version of a text to the most recent version. A decision on the exact semantics to use was left to a
later time, but the two most likely candidates are a date stamp, or a commit hash, or possibly some combi-
nation of the two. It was also left undecided how and at what constitutes the point of publication of a new
version.

Another possibility for the use of the exemplar component would be to use it to identify specific interpre-
tations of an edition, either with or without versioning information included. For example, it could be used
to identify a normalized representation of an edition. As will be seen further below, this might enable us to
use the URNs with subreferences to unambiguously identify individual tokens within the text.

Citation Schemes

At the time this analysis, the New Testament was the only text in the corpus with a traditional canonical cita-
tion scheme of chapter and verse. This is easily represented in the CTS URN structure: *urn:cts:copticLit:nt.1cor.
sahidica:1* identifies chapter one of the SCRIPTORIUM *sahidica* edition of *First Corinthians.*

Many of the other texts in the SCRIPTORIUM corpus tend to have been cited from the modern editions
which compile or reference them much like lost fragmentary texts. As discussed previously, we considered
referencing the individual folios in our citation schemes for the other texts, but that was in conflict with the
concept of a canonical citation per the CTS model, as it cannot carry meaning across different editions of the
abstract work. We chose to postpone declaration of canonical citation schemes for these texts until a point
in time that the corpus was more fully developed.

Even without passage level citation schemes, our URNs *could* be extended with subreferences to identify
individual tokens within our editions:

```
urn:cts:copticLit:shenoute.eagerness.monbgl_29@ⲛⲟⲩⲟⲉⲓⲱ[1]
```

However, there are some significant issues with this approach because it assumes that it is possible to unam-
biguously identify the first instance of ⲛⲟⲩⲟⲉⲓⲱ. Another approach would be to use exemplars to identify
specific interpretations of an edition, as in:

```
urn:cts:copticLit:shenoute.eagerness.monbgl_29.norm20160101@ⲛⲟⲩⲟⲉⲓⲱ[1]
```

---

[21] This digital edition was originally published on Wells' site at http://www.sahidica.org/. It also has been published in the Logos
bible software. The sahidica.org site now no longer exists. The original license and digitization information can be found at http://
copticscriptorium.org/download/corpora/Mark/coptic_nt_sahidic.html.

with the norm21060101 exemplar component used to identify the normalized instance of this text published on January 1, 2016. Such a system would require the numbering of all data tokens to remain stable.

URN Resolution – Enabling Linked Data

The CTS URN identifiers are protocol independent. This is by design in the specification, in order to avoid a dependency on any specific technology. However, this is a drawback if we want our identifiers to be able to participate in a web of linked data served by today's internet technologies. In order to meet our requirements in this regard, we needed to implement a system of resolution for the URNs.

The most straightforward approach, and the one also used by the PDL, was to prefix them with a base URL as a URI space. This also worked nicely with the planned approach to serving different representations as different formats of the same document. The project's portal, http://copticscriptorium.org, links out to all aspects of the project: tools, corpora, documentation, project history, etc. A subdomain (data.copticscriptorium.org) was assigned for the URN resolution service.

Thus, http://data.copticscriptorium.org/urn:cts:copticLit:shenoute.eagerness.monbgl_29 resolves to a landing page for the SCRIPTORIUM edition of Shenoute's *I See Your Eagerness* from folio 29 of the MONB GL codex.

Additionally, http://data.copticscriptorium.org/urn:cts:copticLit:shenoute.eagerness.monbgl_29/norm/ html resolves to the html display of the normalized representation of this text.

We would need to develop a custom resolver to implement this solution, one which should adhere to best practices for linked data. For example, when dealing with versioning requests for the latest version of an identifier may redirect the user via an HTTP 303 "see other" response (Heath and Bizer).

Alternatives to CTS

Having completed the mapping of the SCRIPTORIUM data to CTS identifiers, we still wanted to examine whether other, possibly less complex alternatives, might work equally well. We identified two options for comparison – using more straightforward HTTP URLs or using the Handle system[22].

HTTP URLs

Taking the URL approach, we might simply identify each level of our hierarchy as an individual path component:

http://data.copticscriptorium.org/shenoute to identify the group of Shenoute texts
http://data.copticscriptorium.org/shenoute/A22 to identify the abstract work A22
http://data.copticscriptorium.org/shenoute/A22/MONB.YA to identify the MONB.YA manuscript
http://data.copticscriptorium.org/shenoute/A22/MONB.YA/421-428 to identify folios 421–428

Following RESTful principles[23], resolving each level of the hierarchy would return either the list of items at the next level in the hierarchy or the requested content.

This was appealing because it's very simple and easy for both humans and machines to understand, but it has the drawback of being technology-specific. Although technically http://data.copticscriptorium.org/ shenoute/A22/MONB.YA/421-428 can stand as an identifier even if and when the web is gone, it is more tightly coupled to the technology than URNs are.

Another drawback to the URL approach was that it would be difficult to mint URLs which represented passage level citations in the context of the *abstract work*.

For example, http://data.copticscriptorium.org/nt/mark/sahidica might be the URL identifier for the Sahidica edition of *Mark*, and http://data.copticscriptorium.org/nt/mark/sahidica/1 the URL identifier for Chapter 1 of that edition. It would be undesirable with this structure to use http://data.copticscriptorium. org/nt/mark/1 to identify the abstract notion of Chapter 1 of *Mark* because hierarchically, that puts Chapter 1 of the abtract work at the same level as the entire Sahidica edition of the work.

---

[22] http://handle.net/rfc/rfc3650.html
[23] https://en.wikipedia.org/wiki/Representational_state_transfer

Handles

Handles are another, well-established form of persistent identifier. The following excerpt from RFC 3650 explains:

> "Every handle consists of two parts: its naming authority, otherwise known as its prefix, and a unique local name under the naming authority, otherwise known as its suffix:
>
>   <Handle> ::= <Handle Naming Authority> "/" <Handle Local Name>
>
> The prefixes are assigned to an institution with a local Handle System, the suffixes can be chosen by the local authority. … Local Handle Services are intended to be hosted by organizations with administrative responsibility for handles under certain naming authorities. A Local Handle Service may be responsible for any number of local handle namespaces, each identified by a unique naming authority. The Local Handle Service and its responsible set of local handle namespaces must be registered with the Global Handle Registry. …"

Use of Handles implies an agreement with an organization to provide the Local Handle Service and issue the handles for the data. These agreements can include requirements on the data provider to adhere to specific metadata standards and commit to persistent hosting of the data.

At the current stage of development of the SCRIPTORIUM corpus, we felt Handles were too heavyweight a solution. It was unclear whether we could comply with requirements of existing Local Handle Services and still be able to publish data in an agile manner. And we did not want to take on the responsible for deploying our own Handle service.

We do feel that ultimately it could be desirable to register handles for the SCRIPTORIUM texts and data, and in the future it would be worth exploring ways to do both, for example, by leveraging the URNs as the local name of the handles. CLARIN, a research infrastructure for humanities and linguistic data, provides an analysis of URNs versus Handles in their guide to Persistent Identifiers, and suggests that supporting both approaches is worth considering (CLARIN 2009).

CTS Cost/Benefit Analysis

The decision came down to deciding between CTS URNs, prefixed and served via an HTTP URL prefix, and the simpler RESTful HTTP URLs. The complexities involved in implementing the CTS URN scheme would continue to add overhead to the project in a number of areas. First, in the implementation of a resolution service for the URNs, second in developing new URNs as additional texts are added to the SCRIPTORIUM data set, and third in explaining the choice and expected citation practices clearly to the SCRIPTORIUM users base. An additional argument against use of CTS was that the SCRIPTORIUM could not easily benefit from any existing implementations (at the time of this project) of the CTS API to serve its data. There were a number of reasons for this, the primary one being that the existing implementations expected online editions to be XML documents made up of a single file per edition.

However, these drawbacks were offset by the fact that the level of rigor and intentional decision making involved in thinking through the URN structure gave us increased confidence in the soundness of the identifier structure. CTS URNs also offer at least a theoretical potential for increased data-interoperability across projects, and eventual reuse of implementation code for serving the data via a CTS API. Such interoperability could also be achieved through development, publication and widespread adoption of a common ontology, such as the Linked Ancient World Data Ontology (LAWD)[24], and best practices for its use (assuming data sets are then published which use this ontology and adhere to linked data best practices). However, the issues with representing the hierarchy clearly with only URLs presented a compelling case for CTS, since the paths needed to be both hierarchical and parallel.

---

[24] http://lawd.info

Implementation

Web Service

A web application was developed and deployed on a subdomain of the project's site: data.copticscriptorium.
org. The code is open-source, developed and published on a dedicated public repository at the project's
GitHub organizational account.[25]

Documenting Citation Guidelines

As enabling explicit and stable citation of the texts in various formats was a core requirement per our user
stories, we invested considerable effort in documenting the identifiers and their intended use clearly for end
users of SCRIPTORIUM (**Figure 1**). Each online representation of the texts includes a "Cite this Document"
section that lists the identifiers and how they should be used.

The section also links out to a more detailed page with instructions for how to cite Coptic SCRIPTORIUM
materials, including documents and text visualizations.[26]

Future Work

As discussed above, a number of decisions were postponed to a later phase. These include developing canon-
ical citation schemes for many of the texts and deciding on a definitive approach to versioning and use of
exemplars. We also still need to tackle the challenge of what other distinct data objects in the collection will
be given distinct stable identifiers, and what approach to use for them. This will be necessary to express the
complete data set of SCRIPTORIUM as linked data. As mentioned previously, we will consider adopting the
outputs of the RDA Research Data Collections Working Group for this.

Eventually, we would also like to serve RDF metadata to further enable machine-actionable discovery and
reuse. The following example shows how we might use the LAWD and Open Annotation (OA) ontologies to
do this:

```
@prefix lawd: <http://lawd.info/ontology/>.
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>.
@prefix foaf: <http://xmlns.com/foaf/0.1/>.
@prefix cnt: <http://www.w3.org/2008/content#>.
@prefix perseus: <http://data.perseus.org/texts>.
@prefix script: <http://data.copticscriptorium.org>.
@prefix owl:<http://www.w3.org/TR/2004/REC-owl-semantics-20040210/>.
<script:urn:cts:copticLit:nt.mark.sahidica:1> a lawd:Citation;
      lawd:represents <script:urn:cts:copticLit:nt.mark.sahidica>.
<script:urn:cts:copticLit:nt.mark.sahidica> a lawd:WrittenWork;
      lawd:embodies <script:urn:cts:copticLit:nt:mark>.
<script:urn:cts:copticLit:nt.mark> a lawd:ConceptualWork;
      owl:sameAs <perseus:urn:cts:greekLit:urn:cts:greekLit:tlg0031.
tlg002>.
```

The above statements assert that the resource identified as http://data.copticscriptorium.org/
urn:cts:copticLit:nt:mark.sahidica:1 is a *Citation* which *represents* the *Written Work* resource http://
data.copticscriptorium.org/urn:cts:copticLit:nt:mark.sahidica that *embodies* the *Conceptual Work*
resource http://data.copticscriptorium.org/urn:cts:copticLit:nt:mark.sahidica:1. It further says that this
Conceptual Work resource is the *same as* that identified by the PDL at http://data.perseus.org/texts/
urn:cts:greekLit:tlg0012.tlg001.

Conclusion

The process of understanding the restrictions and intent of the CTS URN specification and applying it to the
SCRIPTORIUM process was labor-intensive at times. We could have applied a more opaque and less mean-
ingful identifier scheme to our data, and still met our need to provide stable, resolvable identifiers for the

---

[25] Luke Hollis of Archimedes Digital and Dave Briccetti developed and implemented the software, available at https://github.com/
CopticScriptorium/cts.

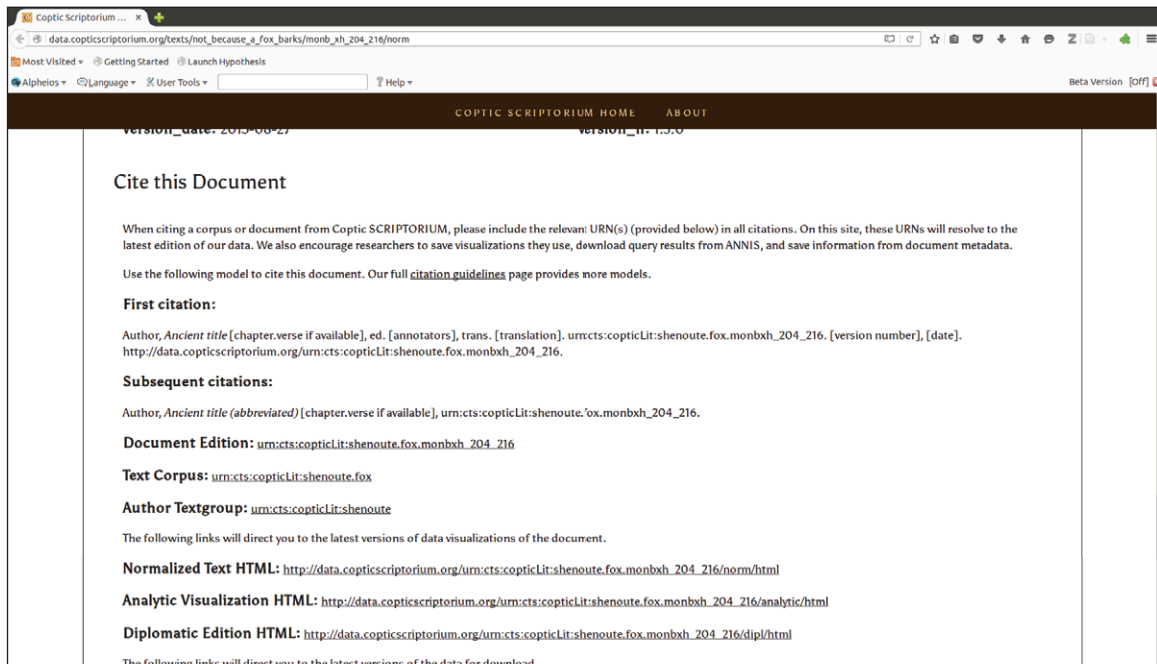[26] http://copticscriptorium.org/citation-guidelines

Figure 1: Cite This Document.

data in the corpus. However, the level of rigor and intentional decision making involved in thinking through the URN structure gave us a greater understanding of our data, an identifier scheme which is meaningful to both humans and machines, and which is not overly dependent any one underlying technology. We foresee that the URNs we have assigned will enable scholars to read, browse, cite and later retrieve digital texts. Our own need to modify an existing system developed by another project (Homer Multitext) rather than import it wholesale illustrates the complexities and particularities of each project's data models and objectives. We hope the documentation of this process in this article will be of use to others seeking models for publishing corpora online as well those seeking pathways for developing models for corpora containing one or more unique characteristics.[27]

Funding Information

Competing Interests
The authors have no competing interests to declare.

References

Almas, B, Babeu, A and **Krohn, A** 2014 Linked Data in the Perseus Digital Library. *ISAW Papers*, 7(3): n. pag.

Amélineau, É 1894 *Histoire Des Monastères de La Basse-Égypte.* Annales Du Musée Guimet 25. Paris: Ernest Leroux.

Berners-Lee, T 2006 Linked Data – Design Issues. N.p.

Chaîne, M 1960 *Le Manuscrit de La Version Copte En Dialecte Sahidique Des Apophthegmata Patrum.* Cairo: Impr. de l'Institut français d'archéologie orientale.

n.d. Checklist of Editions Greek, Latin, Demotic, and Coptic Papyri, Ostraca, and Tablets. N.p., papyri.info.

CLARIN 2009 Persistent Identifier Service. Short Guide. n. pag.

Ehrman, B D and **Pleše, Z** 2011 *The Apocryphal Gospels: Texts and Translations.* New York: Oxford University Press.

---

[27] The process of digitizing and data modeling the pilot corpus underlying this project is documented in the White Paper for the NEH Division of Preservation and Access Grant #PW-51672-14 available at: http://copticscriptorium.org/reports.html and through the NEH public database https://securegrants.neh.gov/publicquery (Schroeder, Zeldes and Platte, 2016).

Elliott, T, Bodard, G, Cayless, H, et. al. 2006–2016 *EpiDoc: Epigraphic Documents in TEI XML* (Online material). Available at: http://epidoc.sf.net.

Emmel, S 2004 *Shenoute's Literary Corpus.* Vol. 599–600. Louvain: Peeters. 2 vols. Corpus Scriptorum Christianorum Orientalium.

Geerard, M, et. al. (eds.) 1974–2003 *Clavis Patrum Graecorum.* Turnhout: Brepols.

Heath, T and **Bizer, C** Linked Data: Evolving the Web into a Global Data Space (1st Edition). *Synthesis Lectures on the Semantic Web: Theory and Technology,* 1(1): 1–136.

Kalvesmaki, J 2014 Canonical References in Electronic Texts: Rationale and Best Practices. *Digital Humanities Quarterly,* 8: n. pag.

Krause, T and **Zeldes, A** 2014 ANNIS3: A New Architecture for Generic Corpus Query and Visualization. *Digital Scholarship in the Humanities*: n. pag.

Lewis, N D 2013 *Introduction to "Gnosticism": Ancient Voices, Christian Worlds.* New York: Oxford University Press.

Renear, A H, Mylonas, E and **Durand, D-G** 1992 Refining Our Notion of What Text Really Is: The Problem of Overlapping Hierarchies. *Research in Humanities Computing,* 4. Selected Papers from the ALLC/ACH Conference, pp. 263–80.

Schirrwagen 2014 Trismegistos. An Interdisciplinary Platform for Ancient World Texts and Related Information. In: Bolikowski, et al. (Eds.) *Theory and Practice of Digital Libraries* Selected Workshops (Communications in Computer and Information Science 416). Cham: Springer, pp. 40–52.

Schroeder, C T and **Zeldes, A** 2016 Raiders of the Lost Corpus. *Digital Humanities Quarterly,* 10(2): n. pag. Retrieved from: http://digitalhumanities.org/dhq/vol/10/2/000247/000247.html.

Schroeder, C T, Zeldes, A and **Platte, E** 2016 *NEH White Paper Report for Coptic SCRIPTORIUM: Digitizing a Corpus for Interdisciplinary Research in Ancient Egyptian.* National Endowment for the Humanities. Retrieved from: https://securegrants.neh.gov/publicquery/Download.aspx?data=EbwGdSyLkD7zoB3W 75cvd%2bXST%2bWypC%2blQBytJB%2bNrYGu%2bmpm3LxKVHtBNsR5g%2fnh8Hvg3LisXiN0b6vo6 0%2f4WP64vQaXK6wNFlnvMz4PpLKidEWo8rM%2bYGAfK%2b9jZIZzSGtKXdlSjTUXCUTpYVLQYQ%3d %3d.

Smith, D N and **Blackwell, C W** 2012 A Gentle Introduction to CTS & CITE URNs. *Homer Multitext Project Documentation.* N.p. Retrieved from: http://www.homermultitext.org/hmt-doc/guides/urn-gentle-intro.html.

Smith, D N and **Blackwell, C W** 2012 Four URLs, Limitless Apps: Separation of Concerns in the Homer Multitext Architecture. *Donum natalicium digitaliter confectum Gregorio Nagy,... The Center of Hellenic Studies of Harvard University*: n. pag.

Zeldes, A and **Schroeder, C T** 2015 Computational Methods for Coptic: Developing and Using Part-of-Speech Tagging for Digital Scholarship in the Humanities. *Digital Scholarship in the Humanities*, 30(suppl 1): i164–i176. DOI: http://dx.doi.org/10.1093/llc/fqv043

Zeldes, A n.d. Coptic SCRIPTORIUM – Lemmatization Guidelines. N.p. Retrieved from: https://github.com/CopticScriptorium/tagger-part-of-speech/raw/master/Coptic%20SCRIPTORIUM%20lemmatization%20guidelines.pdf.

Zeldes, A and **Schroeder, C T** n.d. SCRIPTORIUM Part-of-Speech Tagsets for Sahidic Coptic. N.p. Retrieved from: http://copticscriptorium.org/download/tools/scriptorium_tagset_documentation.pdf.