Interim Report

HG-229371

Koptische/Coptic Electronic Language and Literature International Alliance (KELLIA) Project

Directors:

Caroline T. Schroeder, University of the Pacific

Amir Zeldes, Georgetown University

Institution: University of the Pacific

May 30, 2017

Introduction

This report covers progress on the KELLIA project since November 30, 2016, when the previous interim report was filed. Although this report will make mention of work done by the German partners of the bilateral grant, it will focus on progress made by the US partners. Work on most of the five outcomes outlined in the original grant application is on or ahead of schedule.

US and German KELLIA members Caroline T. Schroeder (University of the Pacific); Amir Zeldes (Georgetown University); Elizabeth Platte (KELLIA Digital Humanities Specialist and Project Manager); and Rebecca Krawiec (Canisius College) will meet at the University of Göttingen in Germany from June 19 to 23, 2017.

Outcome 1 - Milestones for data standards

Outcome 1 is primarily the responsibility of German partners. Uwe Sikora of the University of Göttingen is scheduled to present his recommendations at the KELLIA meeting in June.

Outcome 2 - Server-based batch conversion tools

See Outcome 3 for a report on progress made on the XML converter to convert data output from the Virtual Manuscript Room (VMR) for processing in the Natural Language Processing (NLP) pipeline. In addition, our GitDOX tool (see Outcome 3), an online XML and spreadsheet editor with GitHub-based data storage, implements online server-based conversion of transcription data into the NLP pipeline, as well as conversion of NLP output to our multilayer annotation format.

Outcome 3 - Integration of linguistic tools and methods to produce collaborative digital editions

The November 30 report detailed initial work on an XML converter designed to convert the XML created by transcribing in the Virtual Manuscript Room (VMR) used by German partners into XML structured for the Natural Language Processing pipeline (NLP) created by US partners. This conversion was necessary due to the different data models and encoding standards used by the VMR and the NLP. The converter is adaptable, meaning it can also convert VMR XML into standard outputs, such as the Epidoc subset of TEI. The converter was created by German partner Uwe Sikora in consultation with US partner Elizabeth Platte and has been tested on literary texts transcribed in the VMR with good results. The code will soon be publicly available in the KELLIA GitHub organization repository, and we expect it will be integrated into the VMR as well.

As mentioned in the November 30 interim report, US KELLIA members determined that the VMR could not easily be adapted to the data models and processes used by Coptic SCRIPTORIUM. For this reason, US partner Amir Zeldes and graduate students Shuo Zhang and Emma Manning have created an online XML editor that uses GitHub for data storage known as GitDOX (Git Datastorage Online XML editor), which can be found at https://corpling.uis.georgetown.edu/gitdox/index.py (the site requires an account for access); code is available on Github

(https://github.com/CopticScriptorium/gitdox). This work has also been presented in a conference paper (Zhang & Zeldes 2017, see Publications) and a workshop at the North American Patristics Society Annual Meeting in 2017 (see Events).

Within GitDOX, team members can create and transcribe texts, add metadata, and edit multi-layer annotations (see Outcome 4 for more discussion of the integration of the online multilayer editing tool). This eliminates the need for transcribers, annotators, and editors to install and update software and validation scripts locally. Furthermore, all data from GitDOX can be directly committed to GitHub, meaning that team members no longer have to sync local files to GitHub or merge forks. GitDOX therefore significantly reduces the possibility for human error in corpora development and allows senior editors to control and manage contributions to the project centrally. It also streamlines the process of transcription, annotation, and editing, as texts are available in both transcription and multilayer annotation mode within GitDOX, with uniform automatic validation across data from different users. Future steps, to be completed Summer 2017, include completely linking the NLP pipeline into GitDOX (transcribers can already use the NLP to tokenize texts within GitDOX, i.e. perform automatic morphological analysis of Coptic text) and creating more robust built-in validation. GitDOX is currently being tested by three members of the Coptic SCRIPTORIUM team, and documentation for using GitDOX for the transcribing, annotating, and editing processes in now available on our wiki (http://wiki.copticscriptorium.org/doku.php?id=gitdox_workflow).

Outcome 4 - development of a web-based, multi-layer annotation tool for collaborative text annotations and stand-off markup

The latest version of the web-based spreadsheet program EtherCalc is now integrated into GitDOX, and after updates scheduled for this summer, annotators will be able to seamlessly move from the XML editor through the NLP pipeline and into EtherCalc. Coptic SCRIPTORIUM team members are currently testing the entire GitDOX transcription process, including the integration of EtherCalc. Due to the realtime, online collaborative nature of EtherCalc (similar to Google spreadsheets), the integration of EtherCalc into our workflow allows our Coptic scholars to cooperate and work together on the same documents simultaneously, without risking inconsistencies or version-control conflicts. We view this development as especially game changing for our field, since few departments have multiple Coptic scholars on site, and using these tools we can work as 'virtual teams' in ways that were previously impossible.

Outcome 5 - Sharing, linked data, and textual re-use

The November 30 interim report announced the creation of an online Coptic lexicon with a standalone website created by US partners Amir Zeldes and graduate student Emma Manning (https://corpling.uis.georgetown.edu/coptic-dictionary/) and with full integration into Coptic SCRIPTORIUM's ANNIS interface. Since then, the lexicon has been updated with a "show related items" function. This function allows users to find derivatives and compounds in search results,

giving the online lexicon more of the features that make browsing a paper dictionary and facilitating the discovery of unexpected related information while reading.

In March, we released the second version of the Coptic Treebank (https://corpling.uis.georgetown.edu/coptic-treebank/). The treebank was created by US partner Amir Zeldes and contractor Elizabeth Davidson. The treebank now follows version 2 of the Universal Dependency Guidelines (http://universaldependencies.org/), allowing Coptic to be compared to treebanks in 50 languages, and includes over 8,500 tokens. This growing treebank forms the basis for training stochastic parsers for automatic analysis of Coptic syntax, which is still in very early steps.

US KELLIA partners have also been pursuing opportunities for linking data from digital Coptic projects to other digital projects focused on the ancient world. Since the last interim report, Elizabeth Platte has had conversations with Pleiades (https://pleiades.stoa.org/) and Pelagios Commons (http://commons.pelagios.org/) about linking geographic data from Coptic SCRIPTORIUM's corpora to the Pleiades online gazetteer of the ancient world. Many projects provide link geographic metadata through Pleiades, but our goal is to link ANNIS queries for geographic terms. For this reason, in consultation with Pleiades, we have decided to use Pelagios Commons, which provides infrastructure for linking geographic data, to link our geographic data. Pleiades results include data from Pelagios Commons through an API, allowing our data to be discoverable through Pleiades searches. Platte will prepare RDF files for our data for Pelagios Commons and will document the linking process on the wiki.

As mandated by the original grant proposal, a website for the KELLIA project, created by German partner So Miyagawa in collaboration with US partner Elizabeth Platte, is now publicly available at http://kellia.uni-goettingen.de/. The website is hosted on a server at the University of Göttingen, but the code is available on the KELLIA GitHub organization repository, allowing all organization members to update the website as necessary.

Since the November 30 interim report, we have produced two releases of new corpus data, the first on December 8, 2016 and the second on April 23, 2017. Together, these corpus releases greatly expand the number of published texts. For instance, the Saying of the Desert Fathers corpus now contains 52 published texts (http://data.copticscriptorium.org/urn:cts:copticLit:ap), while the corpus of material from Shenoute's *I See Your Eagerness* (http://data.copticscriptorium.org/filter/corpus=Shenoute.Eagerness) has expanded to over 16,000 words.

Finally, we plan to release a machine-annotated Old Testament corpus shortly. This corpus will complement the full machine-annotated New Testament corpus Coptic SCRIPTORIUM released in 2015 (http://data.copticscriptorium.org/filter/corpus=Sahidica.Nt) and will substantially increase the amount of data available to our users.

Events

On December 9 and 10, 2016, KELLIA and Coptic SCRIPTORIUM members Amir Zeldes, Caroline T. Schroeder and Rebecca Krawiec met at Georgetown University, while Elizabeth Platte and Christine Luckritz Marquis joined the meeting remotely. The team discussed GitDOX, treebanking, the XML converter for the VMR, and linking geographic data in Pleiades, as well as issues with corpora editing and releases. An agenda for this meeting is included as **Appendix 1** to this report.

On March 29 through 31, 2017, Platte met with Zeldes at Georgetown University to learn the publication process for Coptic SCRIPTORIUM texts. Together they worked on the April 23 release.

On May 25, 2017, Schroeder and Krawiec led a digital humanities pre-conference workshop at the North American Patristics Society annual meeting in Chicago. The workshop was titled, "Digital Editions and Text Analysis: Coptic as a Case Study" and introduced participants to tools and processes developed by Coptic SCRIPTORIUM and KELLIA. The agenda included an introduction to digital editions and corpora, working with the online Coptic Dictionary, simple and complex searching Coptic literature in ANNIS, and creating a digital corpus with Epidoc TEI-XML annotations and natural language processing. Detailed tutorials on all these topics are available on GitHub and included as **Appendix 2** to this report (https://github.com/CopticScriptorium/NAPS2017/). Approximately 18 people attended, representing four countries. In terms of stages of their career, participants ranged from graduate students to senior professors and also included librarians. One-third of participants were women and at least two participants were scholars of color. Compare to the approximately 25% of presenters who are typically women at the NAPS annual meeting. (NAPS currently does not track statistics on race/ethnicity; anecdotal evidence suggests the numbers are extremely low.)

US KELLIA members Schroeder, Zeldes, Platte, and Krawiec are preparing to travel to Germany take part in the third KELLIA workshop June 19-23. The workshop will be hosted by the University of Göttingen.

Publications resulting from this project phase

Zhang, Shuo and Zeldes, Amir (2017) "GitDOX: A Linked Version Controlled Online XML Editor for Manuscript Transcription". In: *Proceedings of FLAIRS 2017, Special Track on Natural Language Processing of Ancient and other Low-resource Languages*. Marco Island, FL.

Appendix 1: DC Meeting agenda, December 9 and 10, 2016

In attendance

Amir Zeldes
Caroline Schroeder
Rebecca Krawiec
Elizabeth Platte (remote)
Christine Luckritz Marquis (remote)

Friday, December 9

9:00-10:00 am Check/update Excel plug-ins

10:00-10:30am Firm deadline for Winter/Spring 2017 publication (Beth & Christie will join)

10:30-11:00am Break

11:00am-12:00pm VMR converter (Beth will join)

12:00pm-1:00pm Lunch

1:00-3:00pm Break

3:30-4:30pm Treebanking and general entities (Beth & Christie will join)

4:30-5:30pm linked entities (Beth & Christie will join)

Dinner

Saturday, December 10

9:00-10:00am Further discussion

10:00-11:00am Multilayer editor (Beth will join)

11:00-11:30am Break

11:30am -12:30pm Transcription editor (Beth & Christie will join)

12:30-2pm Lunch

2:00pm-5:00pm Corpus work

Appendix 2: Tutorials for Coptic SCRIPTORIUM Workshop at the North American Patristics Society Meeting, 25 May 2017, Chicago

Readme file

Introduction to Digital Editions and Digital Corpora

Introduction to the Online Coptic Dictionary

Simple and Complex Searches of Coptic Literature in ANNIS

Creating a Digital Corpus

Readme File

NAPS2017

The repository for materials used at the workshop at the 2017 North American Patristics Society annual meeting "Digital Editions and Text Analysis: Coptic as a Case Study."

Authors and Facilitators

Caroline T. Schroeder and Rebecca S. Krawiec

Sources

Many of the tutorials and documentation included here are adapted from existing Coptic SCRIPTORIUM resources listed on our Documentation page. In form, these tutorials are inspired by the University of Victoria Maker Lab's Physical Computing and Fabrication Course at DHSI 2016

License

What are Digital Editions and Digitized Corpora?

What are digital editions and digitized text corpora? This tutorial will generate a discussion about the research purposes of digitized corpora and the how those research methods guide the structure and formation of different corpora.

Explore two digital corpora commonly used by scholars of antiquity and late antiquity

Spend 5 minutes exploring one of the following two digital corpora:

- Perseus Digital Library Greek and Roman Collection
  - Scroll down to Basil for some early Christian texts and click on his Epistles
  - Run a search in the box on the right
  - Click on a word -- what happens?
- Papyri.info text navigator
  - Try filtering using the metadata fields
  - Try changing your font to Coptic and typing in a Coptic word in the search box
  - Try changing your font back to American English and searching for a word in the translation.

As you are exploring, ask yourself:

- What texts are here, what texts aren't here -- what is the corpus?
- What else besides "text" is in this corpus? What other information can I find?
- What kind of research could I/would I do with this text corpus?
- Can I tell how the text has been digitized?

If you have time, check out an aligned digital Hexapla of digitized editions of the New Testament in six languages created by Joel Kalvesmaki as part of his Text Alignment Network project.

Coptic SCRIPTORIUM Digital Editions

Our project publishes digital editions of literary Coptic texts.

Visit data.copticscriptorium.org. (Go directly to this link, or go to our main site at http://copticscriptorium.org and click on Corpora in the menu.) This is our web-based repository designed for reading, browsing, retrieving digital editions of Coptic literary texts. Use the menu to filter for texts.

- Click on Corpus
- Select a corpus, like the Apophthegmata Patrum
  - The "Normalized" button will give you a digital edition of the normalized text.
  - The "Analytic" button will show you an edition of an aligned normalized text, English translation (when available), and part of speech tagging for each word.
  - The "Diplomatic" button will provide a digital edition of a diplomatic manuscript transcription
- If you're looking at any of these digital editions, scroll down to see the document's *metadata* (or information about the document).

Play around with the filtering and browse through the editions. Consider:

- Hover over the text with your cursor. Does anything pop up?
- Can you click on anything? What does it do?
- What texts are here, what texts aren't here -- what are the corpora?
- What else besides "text" is in this corpus? What other information can I find?
- What kind of research could I/would I do with these text corpora?

Want to come back to a text later or cite it? Make a note of its CTS URN. A URN is a uniform resource name (a kind of unique identifier). Our URN's were created using a system called the Canonical Text Services URNs. For example:

- urn:cts:copticLit:ap is the CTS URN for the Coptic Apophthegmata Patrum corpus * urn:cts:copticLit:ap.2 is the CTS URN for the saying in the Coptic AP numbered 2 under Chaîne's numbering system.
- urn:cts:copticLit:ap.2.monbeg is the CTS URN for saying 2 in the Coptic AP as it appears in the manuscript witness known as codex MONB.EG
- You can find the URNs in each document's metadata and in the "Cite this Document" section on the document's page at http://data.copticscriptorium.org
- Try it out: What happens when you type urn:cts:copticLit:shenoute in the box that says "Enter URN" at the top of the screen? What happens when you type urn:cts:copticLit:shenoute.abraham? What happens when you click on this link: http://data.copticscriptorium.org/urn:cts:copticLit:shenoute.abraham ? After you've played around, here's a review key features:

- Normalized editions: English translation (if available) pops up on hover; words linked to the online Coptic Dictionary ("Chapter view" for biblical books.)
- Diplomatic editions: manuscript page number pops up on hover
- Analytic editions are aligned Coptic/English (if available)/Greek (in Mark and 1 Cor corpora)/part of speech)
- You can filter using the menu.
- You can reference and retrieve documents, text groups, and corpora using their CTS URNs.
- All a document's metadata is underneath the edition.
- You can search for a string of characters on any document page using the usual command-f command on your computer
- You can select and copy text, save the html page to your computer, print to pdf just as you do on any other website.

If you want to cite a document or visualization in a publication, see the Citation information after each document's metadata. Also, see our Citation Guidelines page. Always make a note of your document URN and relevant metadata, especially the version number and date of the document. You might also want to save the visualization for your own records by saving the webpage or printing to pdf.

All the editions you see are visualizations generated from text that has been encoded and annotated according to disciplinary standards. The project releases digitized and annotated text in these formats:

- Text Encoding Initiative Extensible Markup Language (TEI-XML) files
- PAULA XML files
- The online installation of the files in the search and visualization tool (or database) we use (called ANNIS)
- The raw files used in the ANNIS installation (relANNIS files)

The following buttons will take you to those data files:

Perseus and Papyri.info also use TEI XML encoding for their digital editions. Coptic Scriptorium is a multi-disciplinary project; we also release our corpora in PAULA XML files, since the PAULA format is used for linguistic research.

If you're creating your own digital corpus, Consider:

- What kinds of things do you want to digitize and why?
- What makes that a "corpus"?
- Do you need to annotate your text in any way? In our corpus architecture, even spelling normalization is an annotation. What annotations do you need? Why and how?
- What kind of metadata do you need to provide?
- What kind of access will you allow others to have to your corpus? (Consider a license for your corpus and data.)

If you're interested in aligning texts and translations, here are some tools:

- Alpheios Text Alignment App (allows you to export XML files of your alignment; created in collaboration with Perseus)
- Text Alignment Network
- Ugarit text alignment

If you're interested in Coptic New Testament and Old Testament editions, check out these projects in Germany:

- New Testament Virtual Manuscript Room which includes Coptic
- Coptic Old Testament Project

Coptic Dictionary Online

This tutorial will introduce you to the basic functionality of the Coptic Dictionary Online, created by the KELLIA project. The KELLIA project is a collaboration between German and US researchers working in Coptic Digital Humanities, funded by the National Endowment for the Humanities and the Deutsche Forschungsgemeinschaft. (more information)

First simply type the search window a Coptic word (in utf-8/unicode Coptic characters). You can control for dialect and part of speech (N=noun, V=Verb, etc.) or you can leave those options blank.

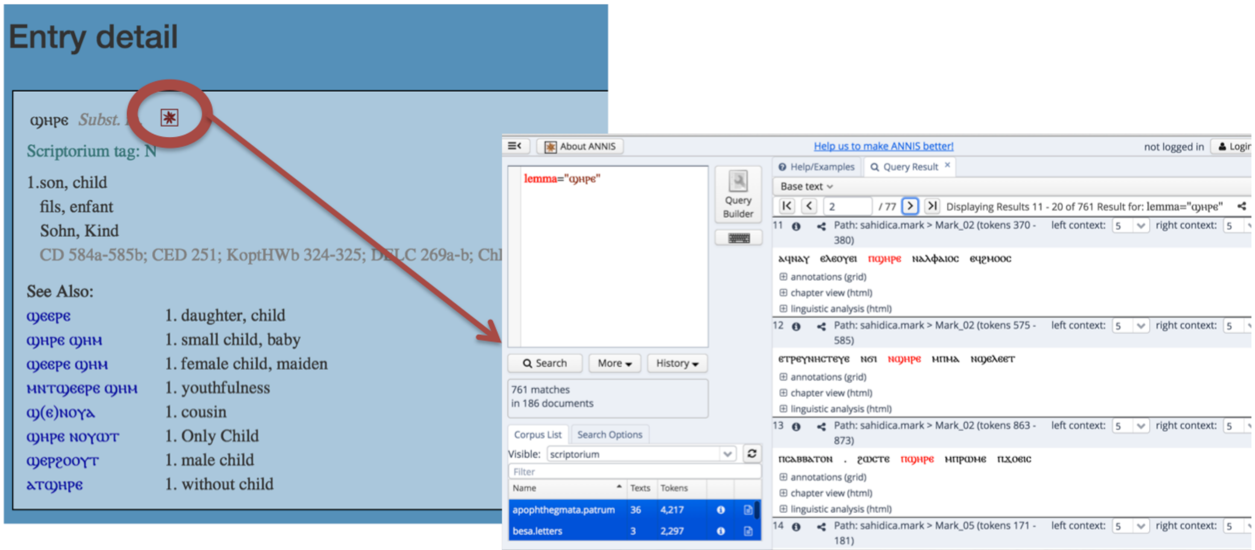
Be sure to check out the search tips on the Help page. Play around with searches for words containing particular suffixes, prefixes, or character strings.

You can even search translations. Forget what the Coptic word for "river" is?

- Try typing "river" into the quick search bar on the top of the screen
- Or go to the Dictionary home page and try typing "river" into the Translation box.
- You can search translations in English, French, and German

Try searching for a word, such as ϣⲏⲣⲉ ("child, son").

- Click on the word ϣⲏⲣⲉ to see the entry
- Click on the little plant icon. This is the icon for the search tool for our digital corpus (ANNIS). You will get the results for all the matching lemmas for ϣⲏⲣⲉ in our corpora. Take a look.


Entry detail
ϣⲏⲣⲉ Subst.
Scriptorium tag: N
1.son, child
fils, enfant
Sohn, Kind
CD 584a-585b; CED 251; KoptHWb 324-325; DELC 269a-b;
See Also:
ϣⲉⲉⲣⲉ 1. daughter, child
ϣⲏⲣⲉ ϣⲏⲙ 1. small child, baby
ϣⲉⲉⲣⲉ ϣⲏⲙ 1. female child, maiden
ⲙⲛⲧϣⲉⲉⲣⲉ ϣⲏⲙ 1. youthfulness
ϣ(ⲉ)ⲛⲟⲩⲁ 1. cousin
ϣⲏⲣⲉ ⲛⲟⲩⲱⲧ 1. Only Child
ϣⲉⲣϩⲟⲟⲩⲧ 1. male child
ⲁⲧϣⲏⲣⲉ 1. without child
About ANNIS
Help us to make ANNIS better!
not logged in
lemma="ϣⲏⲣⲉ"
Query Builder
Search
More
History
761 matches
in 186 documents
Corpus List
Search Options
Visible: scriptorium
Filter
Name
Texts
Tokens
apophthegmata.patrum 36 4,217
besa.letters 3 2,297
Help/Examples
Query Result
Base text
/ 77
Displaying Results 11 - 20 of 761 Result for: lemma="ϣⲏⲣⲉ"
Path: sahidica.mark > Mark_02 (tokens 370 - 380)
left context: 5
right context: 5
annotations (grid)
chapter view (html)
linguistic analysis (html)
Path: sahidica.mark > Mark_02 (tokens 575 - 585)
Path: sahidica.mark > Mark_02 (tokens 863 - 873)
Path: sahidica.mark > Mark_05 (tokens 171 - 181)

ANNIS Tutorial

Introduction to ANNIS

Coptic Scriptorium uses the ANNIS search and visualization tool. You can access Coptic Scriptorium's corpora in ANNIS in multiple ways:

- Go directly to https://corpling.uis.georgetown.edu/annis/scriptorium and run a query
- In the Coptic online dictionary, search for a word; click on the ANNIS icon to find instances of that lemma in ANNIS
- When browsing documents at our portal for reading Coptic texts, click on the "Search ANNIS" button to query that corpus.

This tutorial will:

- introduce you to our documents and corpora in ANNIS
- show you how to perform simple word searches
- show you how to perform more complex searches for annotations and metadata
- demonstrate how to generate word frequency lists
- show you how to download search/query results
- provide models for citing or linking to the data in your publications

ANNIS Corpus Browser

When you arrive at https://corpling.uis.georgetown.edu/annis/scriptorium, you will see the list of publicly available corpora on the lower left of your screen. (On the right, you will see a list of sample queries for our corpora -- more on that in a minute.)

Each *corpus* contains multiple *documents*.

- Each corpus has its own *metadata* (information about the corpus, such as all the editors/annotators who worked on the corpus, license information, the date this version of the corpus was released, etc.)
- Each document within the corpus also has its own *metadata* (tite of the document, manuscript information if the text is from a manuscript, specific

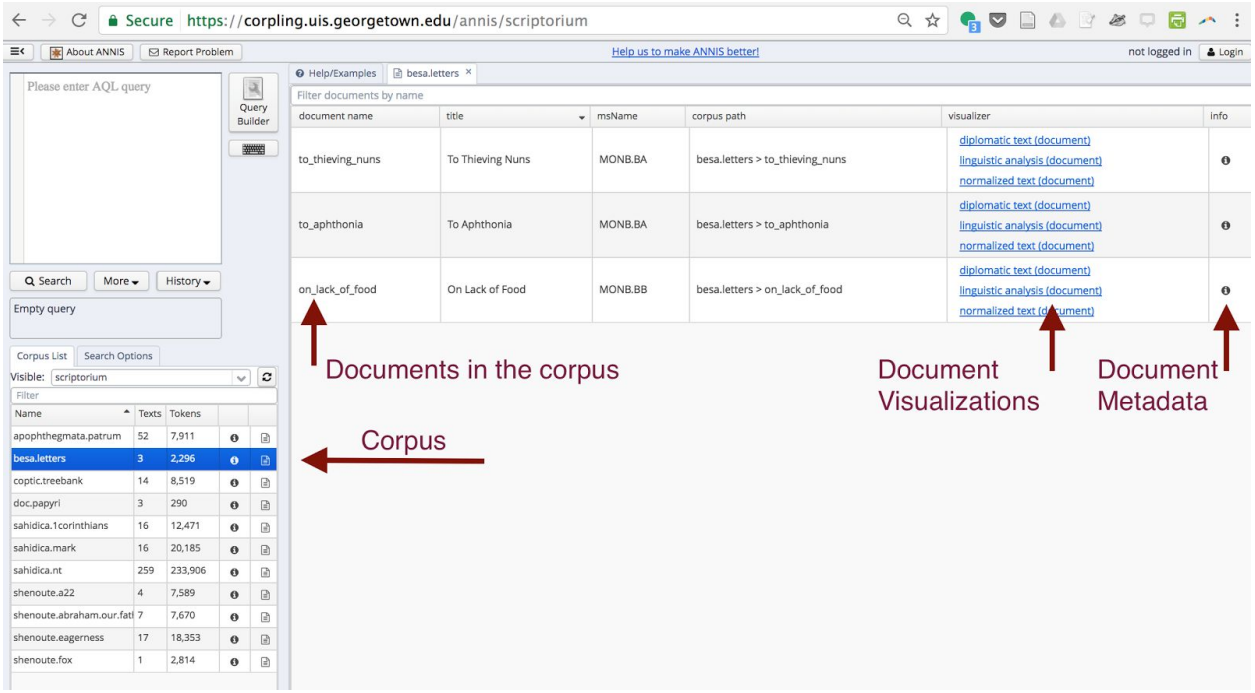
editors/annotators for that document, translators, the date this version of the document was released, etc.)

Look at the list of corpora:

1. To find out more information about any corpus, click the "i" information button for that corpus. A window will appear with:
- a dropdown menu at the top listing all the documents in a corpus
- on the left, *metadata* (information about the corpus, such as annotators, translators, the date this version of the corpus was released)
- on the right, all the annotations available for texts in this corpus. These annotations will make more sense to you after you run a few searches. (We also have documentation on our wiki if you're really interested.)

1. To see a list of documents in any corpus, click the document icon ⌄ for that corpus name
- Click the "i" information button for any document for more information
- You'll also see a list of visualizations for each document (the same visualizations available at http://data.copticscriptorium.org)

➡Try it: What happens when you click on a link for a visualization?



1. You can filter the list of corpora also:

➡Try it: What happens when you type in "shenoute." (without the quotation marks) in the Filter box above the list of corpora?

➡What happens when you click the ⏝ button ?

Basic Search

Let's start by using the example queries provided for any given corpus.

➡1. Try it: Click on the first corpus, apophthegmata.patrum. Then play around with the sample queries or follow the following steps

- In the panel on the right, click on the query to search for the word "ⲁⲡⲁ". Note: this searches for the *normalized* word, meaning spelling variants have been normalized, diacritics removed, and missing/damaged letters reconstructed.
- In the search results on the right, your query term should appear in red (possibly within a Coptic bound group).
- See the phrase "Base text" at the top of the list of results?
  - Change the base text from "norm_group" to "norm"? How does this change how the results appear on the screen?
  - Change the base text to "orig"? (Note: orig is an abbreviation for "original text" transcription)
- You should see the same visualizations we've seen before (Analytic, Diplomatic, Normalized views). Click on the + next to "analytic view".
  - Can you see your search result in red again?
  - This view visualizes three annotations of the textual data: part of speech annotations, the normalized Coptic text, and an English translation.
  - Check out the other two visualizations. What information is available?
- What happens when you click the "i" information icon for the first search result? What information does this give you?
- To view ALL the annotations for any given query result, click on "annotations (grid)".
  - All annotations for that stretch of text will appear as layers below.
  - Some annotations have been manually encoded; others have been added using our Natural Language Processing tools

ANNIS uses a multi-layer annotation model, where a base text appears followed by layers of annotations on that base text. You can have any number of customized annotations. All our paleographic and manuscript annotations (lacunae, page breaks,

column breaks) follow a set of annotation and encoding standards known as the TEI-XML (the Text Encoding Initiative standards for extensible markup language). Specifically, we use the Epidoc subset of TEI XML, the same encoding standards that Papyri.info uses.

➡2. Try it: Let's create your own simple searches for words.

- Modify the search we just did by typing your own favorite Coptic word where "ⲁⲡⲁ" appears. Click "Search". Check out the results.
- Don't have a Coptic font installed on your computer? Click on the little keyboard to the right of the search pane!
- Let's now search for your favorite word in more than one corpus. Control-click on a Mac/right-click on a PC on another corpus name in the corpus list in the lower left. Click Search.

➡3. Try it: Create simple queries for information other than words.

- Search for norm="ⲥⲟⲛ" in your chosen corpora
- Now search for lemma="ⲥⲟⲛ". What's the difference in the results?
- Search for all Greek words in Shenoute's "I See Your Eagerness": click on tge shenoute.eagerness corpus and search for lang="Greek"(link)
- Search for all words with the morpheme "ϩⲛⲧ" in Shenoute's "Not Because a Fox Barks": click on the shenoute.fox corpus and search for morph="ϩⲛⲧ" (link)
- Search for all proper names in Warren Wells' Sahidica edition of the Gospel of Mark: click on the sahidica.mark corpus and search for pos="NPROP"(link)
- Play around with some simple searches.

➡4. Try it: You can click on the History button to see all the previous queries you've run in your current ANNIS session.

Complex Searches

You can also use regular expressions and the Annis Query Language to create complex queries, searches for sequences of characters, queries for two or more annotations, etc.

➡5. Try it: Select a corpus, like 1 Corinthians, and try the following queries. (Type or cut-and-paste.) What kind of results do you get?

- norm_group=/ⲡⲉⲧ.*/

- norm=/.*ⲟⲥ/
- norm=/ⲥ[ⲟⲱ]ⲧⲏ/

Hint: the .* in the query syntax signals that you want to search for any character(s).

You can also search within a translation, if your corpus has a translation. (Not all do.)

➡6. Try it: Select the 1 Corinthian corpus. Try the following queries. What's the difference?

```
translation=/.*brother.*/
translation=/.*[Bb]rother.*/
```

You can search more than one field at the same time.

➡7. Try it: Say you're interested in proper names. Select the corpus for Abraham Our Father. Compare the following queries

- `pos="NPROP" _o_ lang="Greek"`
- `pos="NPROP" _o_ lang="Hebrew"`
- `pos="NPROP" _o_ lang=/.*/`

Note: We tag loan words for language of origin based on the oldest possible language. To find all loan words, use the lang=/.*/ query.

You can also add metadata to your queries.

➡8. Try it: Select the Abraham Our Father corpus. Search for all the appearances of "ϣⲉⲉⲣⲉ" in the codex MONB.YA: norm="ϣⲉⲉⲣⲉ" & meta::msName="MONB.YA"

- Play around with other metadata fields. To find all words in documents edited by Rebecca S. Krawiec, select your corpora and search:

```
norm & meta::annotation=/.*Krawiec.*/
```

There's lots of fun stuff you can do with regular expressions and the ANNIS Query Language:

- Find either circumstantial converters or focalizing converters: pos=/CCIRC|CFOC/
- Find either form of the same verb: norm=/ⲥ[ⲟⲱ]ⲧⲏ/
- Query for things following each other: To search for a copular pron sentence (a copula following a pronoun): pos="PPERI" . pos="COP"
- Query for nearness: To find "daughter" within 50 tokens after "son": norm="ϣϩⲣⲉ" ^* norm="ϣⲉⲉⲣⲉ"

Know your corpus and annotations when doing research. For example, in our corpus, a compound word containing both Greek and Coptic contains a language tag only for the Greek morph within the compound. (E.g., in ⲣⲭⲣⲉⲓⲁ, only ⲭⲣⲉⲓⲁ receives the Greek tag. Hence, we use syntax for finding overlapping search fields ("o") rather than equivalent fields ("="). `Lang="Greek" _=_ pos="V"` (link) finds all verbs that are Greek loanwords; `Lang="Greek" _o_ pos="V"` (link) finds all verbs that are Greek loanwords or are compound words with Greek loan words as part of the compound. Compare the results in the links.

Word Frequencies

ANNIS allows you to find word frequency lists for our corpora.

➡1. Try it: Select the shenoute.eagerness corpus.

- type in the following query to find all the words in the corpus: norm
- Below the query window, you should see a button for "More." Click on it and select "Frequency Analysis." Click "Perform Frequency Analysis"
- Both a chart and a list of word frequencies will appear.
- You can see your frequencies on a linear scale or a log scale

➡2. Try it: Download your frequency list by clicking the "Download as CSV" button

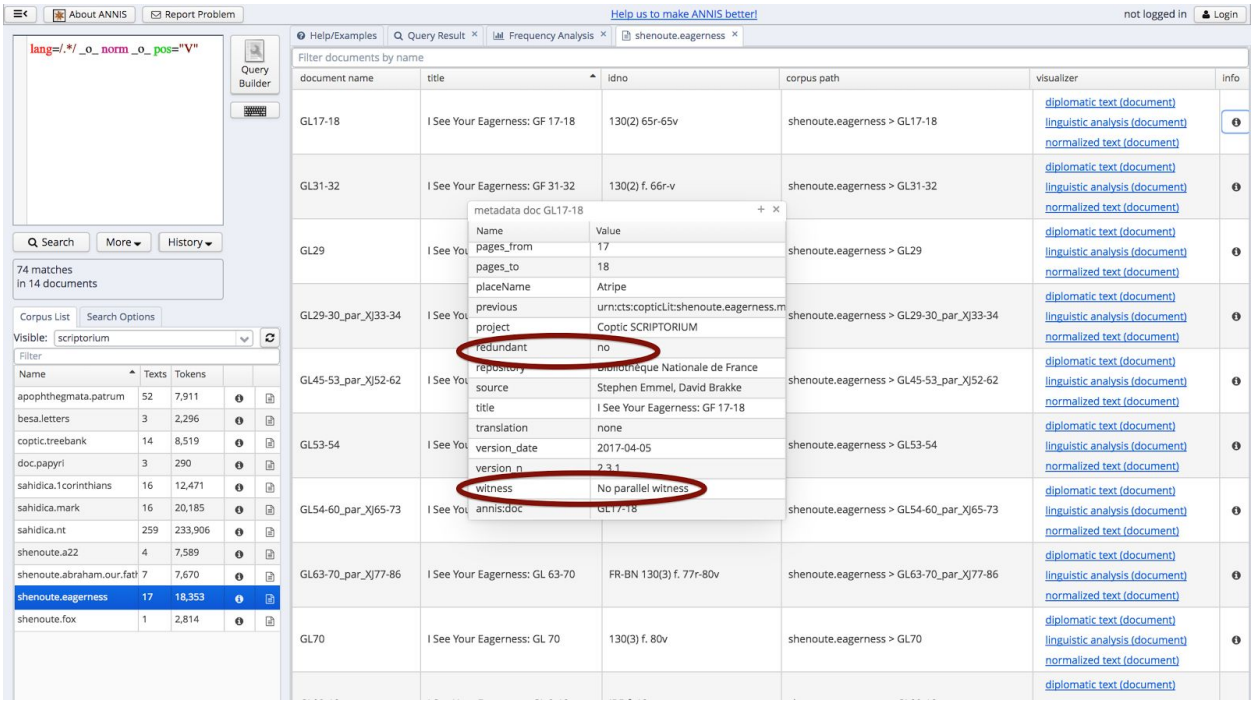
You can also produce frequencies for more refined lists. *Be sure to close the "Frequency Analysis" pane to clear your data before you start a new analysis.*

➡3. Try it: Create lists for loan words in our corpus.

Remember: If you have just run a frequency analysis, then close the current "Frequency Analysis" pane first. Do this (or click "new analysis") between each new frequency analysis.

A. Find the Greek loan words in the shenoute.eagerness corpus using this query: `lang="Greek" _o_ norm` * Enter the query. (If you've closed the Frequency Analysis pane, click "More" then "Frequency Analysis") * Delete all rows EXCEPT "norm" (since you want the frequency of each normalized word) * Click "Perform Frequency Analysis" B. Can you do the same to find *all* loan words in the shenoute.eagerness corpus (remember to hit "new analysis" first): `lang=/.*/ _o_ norm` C. Can you do the same to find all loan words that are *verbs* (remember to hit "new analysis" first): `lang=/.*/ _o_ norm _o_ pos="V"`

BONUS question: What do you do about corpora that contain more than one manuscript witness to the same text? I See Your Eagerness is one such corpus. In some places, we have parallel manuscript witnesses to the same text. So if you run a straight word frequency list, you'll get duplicate "hits". For this corpus (and future versions of other corpora) we encode parallel witnesses in the metadata fields. When you click on the "i" information button for a document, you'll see metadata fields for "witness" and "redundant".



➡4. Try it: Run a freqency analysis using the following query: `lang=/.*/ _o_ norm _o_ pos="V" & meta::redundant="no"`

- Remember to click "new analysis" to clear your old frequency data first!
- Remember to delete rows for everything except norm when you run the analysis.

- Are the results *different* from the results from #3 above?

Again: know your corpus so you understand the numbers. Spend some time looking at the metadata, understanding the annotation layers, and running queries to see how the annotations and textual data work. In our corpora, we designate as redundant the withness(es) with the most damage or lacunae.

Download Your Results

We encourage all researcher to keep records of their research in ANNIS. This includes queries, the corpora on which the queries are run , the version number and version date of the corpora, and the results.

There are multiple ways you can download the results of your query by clicking More > Export underneath the query panel. Each way or format works well for a different discipline or research objective. For most people who work with texts as philologists, historians, or religious studies scholars, we recommend using the GridExporter. The GridExporter allows you to tell ANNIS which annotations and which metadata you want to export.

➡Try it: Run a query (such as this one) and download your results.

- Run the query
- Click "More" > "Export"
- In the Exporter dropdown menu select GridExporter
- In the "annotation keys" box, type the annotations you want to export. Try: orig, norm, translation, pb_xml_id to export the original manuscript text, normalized text, translation (if available), and the page number of the manuscript
- In the Parameters box type numbers=false;metakeys=title,version_n,version_date to export the document title, version number, and version date for EVERY hit in your search.
- Click Perform Export
- Click Download
- You can open this text file in any text editor (such as TextEdit, Text Wrangler, etc.)
- If you want more annotations (such as part of speech tags) add them to the "annotation keys" box; be sure to use the correct name for the annotation

- If you want more metadata (such as the names of editors or translators of each document), add them to the "Parameters" box; be sure to use the correct name for the metadata field

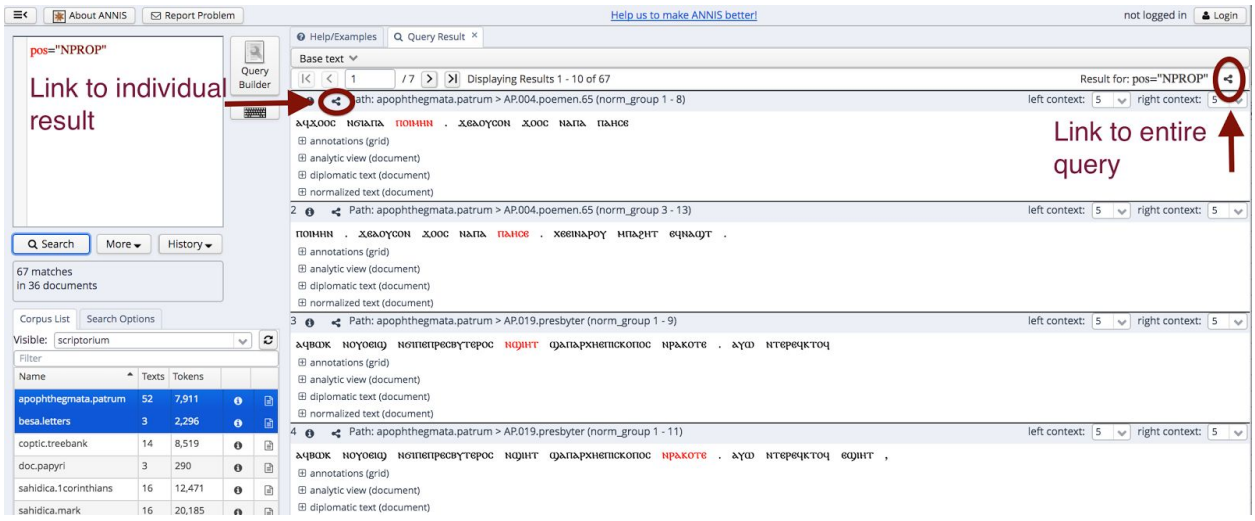
Citing and Linking to Your Data

When researching our corpora for a future publication, please note the date and version number of the documents or corpora while you are conducting your research. (This information is in the corpus and document metadata accessed via the information button(s) for each corpus and each document within a corpus.) We update our corpora regularly and recommend all citations include the version number and date of the resources used, as described below. (If you conducted research in the past and did not note the version and date of the corpus at that time, then please cite the date you accessed the corpus.)

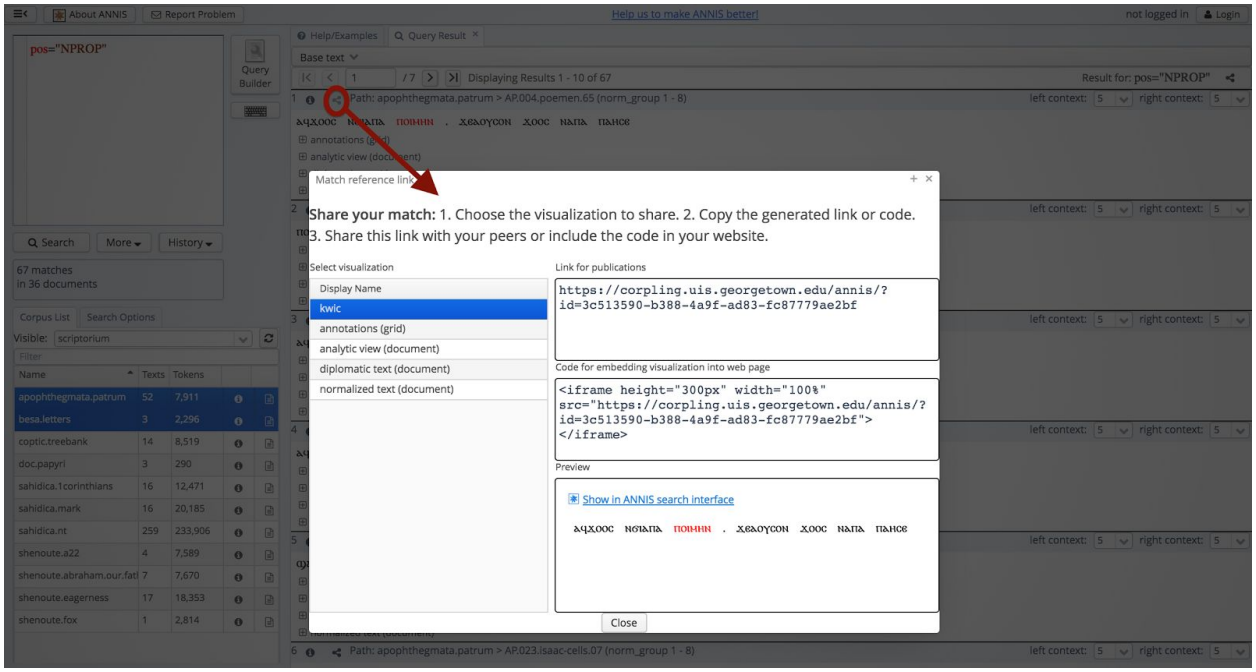
We have Citation Guidelines with examples for how to cite the project, the project site, individual corpora, and individual documents in your bibliography and footnotes. If you are using documents or queries on only one corpus, then you may cite only that corpus.

When citing more than one corpus, we recommend citing the corpora and versions of each corpus used.

You can save a link to a query or even to a query result.



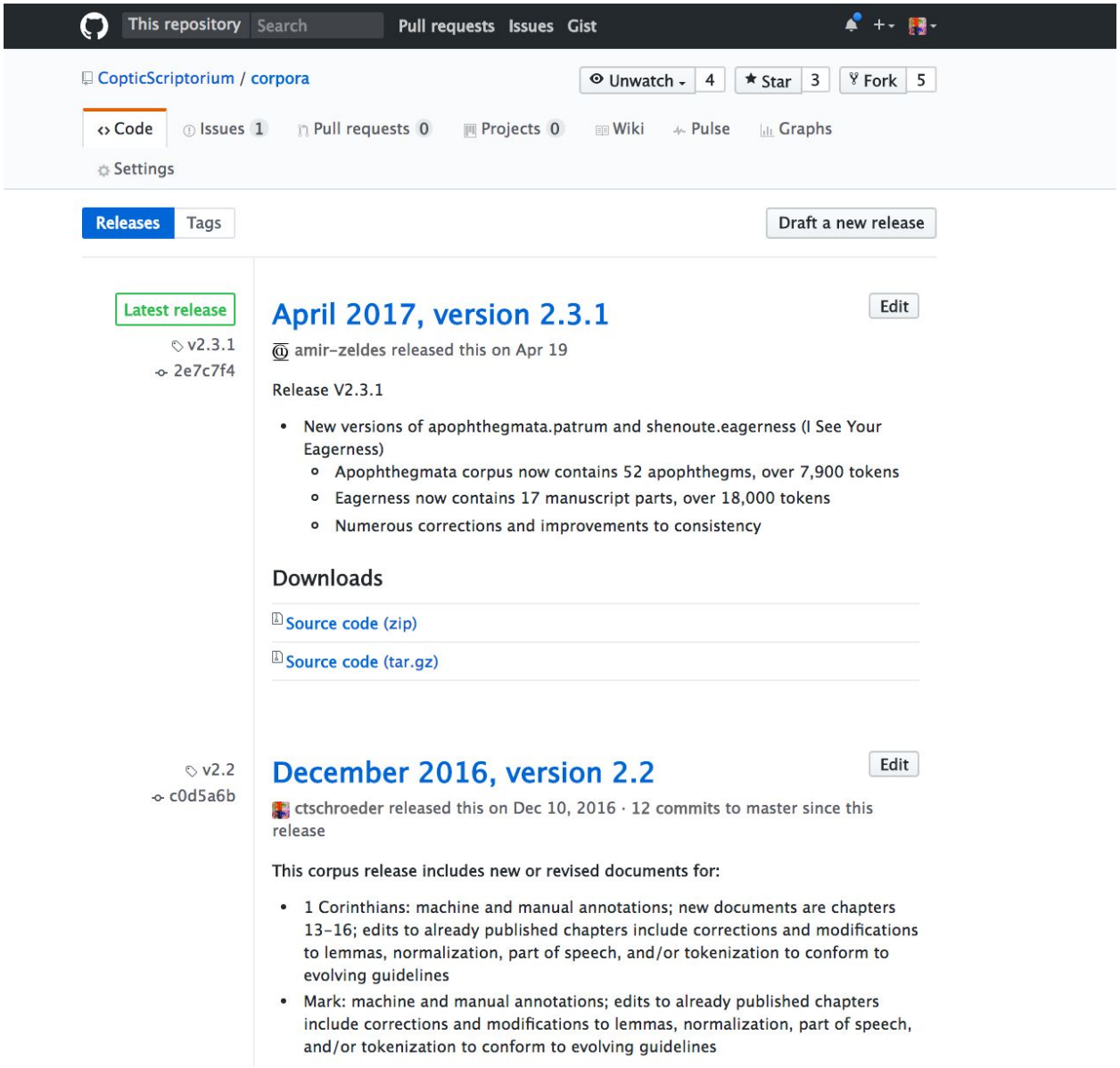
If you want to embed a result in a blog, webpage, or other electronic publication, you can do that too!



Some DH researchers recommend providing access to your data when you publish your analysis. You can do this in a number of ways:

1. Link to our project's raw data on our GitHub corpus repository.

- Link to the version that corresponds to your data. See our release page.



- So, for research conducted in May 2017, link to the April 2017 version. For research that was conducted in January 2017, you would want to link to the December 2016 version

1. Download your query results using the process described above and post them on your own site; link to them in your publication.
2. Link to your query on our ANNIS site.

Important note: the URLs for the query and result links are stable, but the core text data may change if we update the corpus or documents you are querying. We update regularly to add more documents to a corpus, to add new annotations, or to make

corrections. We encourage all researchers to download query results and cite the version number(s)and date(s) of the data used.

Digitize and Annotate a Text Using Coptic Scriptorium's GitDox Tool and Natural Language Processing Tools

In this tutorial, you will learn how to digitize and annotate a short Coptic text using our GitDox tool and NLP tools. For participants in the NAPS workshop, we will provide a text and all the other necessary materials. Check with the facilitators. In groups of 2-4 people, you will transcribe a text and provide basic paleographic/manuscript encoding.

For more detailed documentation, visit our wiki; some of this tutorial is taken from that documentation.

If you want to digitize a specific text for inclusion in the Coptic Scriptorium corpora and database, we would like to collaborate with you. Please contact us!

Login and Orientation to GitDox

GitDox is an online XML and spreadsheet editor which uses GitHub for data storage and versioning. Coptic Scriptorium currently uses GitDox to transcribe texts and edit them in a spreadsheet.Although you don't need to have a GitHub account to use GitDox, it's helpful. For the purposes of the NAPS workshop, you don't need one. If you want to collaborate with us or use GitDox in the future, you can get an account later.

GitDox is located at https://corpling.uis.georgetown.edu/gitdox/ . Navigate there on your computer, and login using the username and password we provided you at the workshop.

When you log in to GitDox, you see a list of current documents. Use the dropdown menu above the list to display only documents from a certain corpus. You can also use the arrows to the right of each column name to sort by that column.

Documents are assigned to users (as noted in the fifth column). Please only edit documents assigned to you. If you believe a document should be assigned to you but isn't, please contact the person to whom it is currently assigned to confirm that you should be assigned the document before editing it.

Find the text we assigned to you in the workshop. Click on the "Edit" button for your text. Assign it to yourself.

Saving and Committing

The transcription mode of GitDox has two options for saving work: save and commit. The save button saves changes within GitDox but does not commit those changes to GitHub. Your permission levels will not allow you to commit to GitHub; one of the workshop facilitators will complete this step of the process. Because GitDox depends on an internet connection, it's a good idea to save your work frequently while you're working using the save button.

Adding metadata

Use the button at the bottom of the page to add metadata. We've added some of the important metadata for th etutorial. You will need to add your own names. Click the metadata button. Enter "annotation" in the first window and your group's names separated by commas in the second (e.g., Caroline T. Schroeder, Rebecca S. Krawiec). After you click submit, the metadata will appear in a chart on the bottom of the page.

You can't edit metadata you've already submitted, but you can delete the entry from the list and re-enter the correct field and information.

Transcribing

Now begin typing in your text. Use a utf-8 (Unicode) characterset, such as the Antinoou font and keyboard. Transcribe as you would any manuscript. Use regular returns to create line breaks

You'll notice two "tags" already in your otherwise empty document. These are XML tags for annotating text. If you want to annotate something in your text, you'll wrap the relevant text in "tags". The first tag (the open tag) is in brackets, such as `<TEI>`; the close tage has brackets and a slash, such as `</TEI>` (If you want, you can read more about TEI XML and the Epidoc subset of XML.)

Start your transcription *inside* the TEI tags.

You will want to encode for information about pages and columns.

- `<pb></pb>` The page break tags will wrap around all the text in a given page. We use what's called an attribute to encode the page number of the manuscript.
- `<cb></cb>` Column break tags will wrap around all the text in a given column.

When you open an angle bracket, GitDox suggests tags that are currently available. GitDox will also suggest attributes for tags. Improperly closed tags and other errors are highlighted in red.

If you need to provide other information, such as if a letter in the manuscript is large or ekthetic (hangs out in the margin), see the cheatsheet provided in the workshop or our longer Transcription Guidelines for instructions.

Below, see a sample text encoded as an example:

```
<TEI>
<pb xml:id="YB307"><cb><note note="large embellishment in margin before the first letter of the page andpage number ⲧⲍ visible in upper right corner of page"><hi rend="ekthetic">Ⲉ</hi></note>ⲧⲃⲉⲡⲁïⲙ̣ⲛ̣
ⲟⲩⲥⲓⲁϭⲉⲛⲁ
ϣⲱϫ̄ⲛ̄ϩⲁⲛ̣<hi rend="small">ⲉ</hi>ϥ
ⲛⲟⲃⲉϫⲉⲛ̄ⲛ̣ⲉϥ
ⲙ̣ⲉⲧⲁⲛⲟⲉⲓ
ⲁ̣ϣⲧⲉⲧⲙⲉⲧⲁ</cb>
<cb>ⲛⲟⲓⲁⲛ̄ⲥⲁⲧⲣⲉⲛ
ⲥⲁϩⲱⲛⲉⲃⲟⲗ
ⲛ̄ⲕⲣⲟϥⲛⲓⲙϩⲓ
ⲙⲛ̄ⲧ̄ⲁⲧⲥⲱ̄ⲧ̄ⲙ̄
ⲛⲓⲙⲧⲛⲟⲩⲛⲉ
ⲛϩⲱⲃⲛⲓⲙⲉϥ</cb></pb>
</TEI>
```

Either while you are typing or after you are done transcribing, separate Coptic bound groups with a space or underscore. For example:

ⲁϥⲥⲱⲧⲙ̄ ⲛ̄ ϭⲓⲡϫⲟⲉⲓⲥ ⇒ ⲁϥⲥⲱⲧⲙ̄ _ⲛ̄ ϭⲓⲡϫⲟⲉⲓⲥ_

Be sure to separate the punctuation, as well.

ⲁϥⲥⲱⲧⲙ̄ ⲛ̄ ϭⲓⲡϫⲟⲉⲓⲥ· ⇒ ⲁϥⲥⲱⲧⲙ̄ _ⲛ̄ ϭⲓⲡϫⲟⲉⲓⲥ_·_

Put a separator between all bound groups, even if a bound group ends at the end of a line:

ⲁⲩⲥⲱⲧⲙ̄ ⲁⲩⲥⲱⲧⲙ̄_ ⲛ̄ⲃⲓⲡⲭⲟⲉⲓⲥ ⇒ ⲛ̄ⲃⲓⲡⲭⲟⲉⲓⲥ_·_

If you need to provide other information, such as if a letter in the manuscript is large or ekthetic (hangs out in the margin), see the cheatsheet provided in the workshop or our longer Transcription Guidelines for instructions.

When you are done typing in your text, be sure to SAVE.

Tokenizing

Our natural language processing tools will automatically divide bound groups into words. This process is called tokenization. Click the NLP button under the transcription box to automatically tokenize your transcribed text. Correct the tokenization in GitDox by adding pipes between the words.

Processing using the NLP

We expect an automated link between the XML editor, the NLP, and an online spreadsheet editor to be available in summer 2017. However, as of May 2017, the following steps are necessary to move transcribed texts from the XML editor through the natural language processing tools.

After checking tokenization and saving the final text, copy the text from GitDox and paste it into the online NLP pipeline at https://corpling.uis.georgetown.edu/coptic-nlp/. Make sure you select “from pipes in input” under tokenize. Process the text, copy the resulting SGML, and paste it into a new file (you may use a text editor such as text wrangler or notepad) on your computer. Save the file on your computer.

In your file on GitDox, use the mode drop-down menu to change to the spreadsheet mode. A blank spreadsheet should appear. Scroll to the bottom of the page and use the

upload function to upload the file you saved on your computer. The spreadsheet should be populated with the information from your file.

Editing the annotations

You should now see annotations in layers, very similar to the grid annotations we saw in the ANNIS database.

Check the annotations! You probably don't know all our part of speech abbreviations, but you can tell if the NLP tools caught all the loan words, or if they expanded nomina sacra in the normalized layers.

(Handout) Coptic SCRIPTORIUM Cheat sheet for commonly used EpiDoc XML tags for manuscript transcription

Transcription for a digital corpus combines traditional transcription methods with standardized coding, called “tagsets.” This will produce a document that the Natural Language Processor (NLP) on CopticScriptorium.org will be able to process and annotate to normalize spelling, diacritics and tag for part of speech, loan words, and lemmas.

When you transcribe, you will be adding your own annotations for manuscript information (column breaks, page breaks and page numbers, ornamented characters

A full description of our diplomatic transcription guidelines is available on the website. For the purposes of this workshop, we have prioritized the most common and useful tagsets to include in our on-line transcription editor.

IMPORTANT: All tagsets must have at least one letter or punctuation mark between the tags. All tagsets should be nested (e.g., <pb><cb>columnofCoptictext</cb></pb> not <cb><pb>columnofCoptictext</cb></pb>)

Page breaks <pb> at the beginning of the page and </pb> to close the page

Column breaks <cb n=”1”> to begin the first column and </cb> to end the column; use <cb n=”2”></cb> for the second column

Character ornaments:
Use <hi> </hi> (for “highlighting”) with an attribute added to specify the kind of highlighting, usually rend (for rendering). E.g. type <hi rend=”ekthetic”></hi> wrapped around the ekthetic charact(ers). Transcriptions can combine two attributes which should be entered with a space, no punctuation, between them.

<hi rend=”[attribute]”></hi>

1. Outside the left margin: rend=”ekethetic”
2. Large letters: hi rend=”large”
2. Ekthetic & large: hi rend=”ekthetic large”
3. Letters that are above the line: rend=”superscript”
4. Letters that are below the line: rend=”subscript”
5. Letters that stretch tall above the line: rend=”tall”
6. Letters that stretch long below the line: rend=”long”
7. Letters that appear in red ink: rend=”red” (or brown, green, etc.)

Additional notations: Use a note for decorations in the margins or other information that might be difficult to tag. Wrap the tags around the nearest text.

<note note="there is a drawing of a bird below the column"></note>

Lacunae, damage:

If you find lacunae or damaged characters, transcribe them using the usual conventions (underdot for partially visible/ambiguous character, square brackets for missing and reconstructed characters, etc.) Here are some common tags (taken from the papyri.info cheat sheet):

Letters ambiguous outside of their context:	ⲁⲭⲥⲫⲧⲏ (no tags, underdots only)
Vestiges of letters visible but illegible	<gap reason="illegible" quantity="3" unit="character"/>[...]</gap>
Characters lost but restored	<supplied reason="lost">[ⲁⲭⲥ]</supplied>ⲱⲧⲏ
Characters lost (lacuna such as ms tear)	<gap reason="lost" quantity="3" unit="character">[...]</gap>

Note: for papyri.info editors, we use very similar tags and conventions, but there are some differences due to our multilayer annotation format.