Grammateus: from Ancient Greek Papyri to a Web Application

Elisa Nury

Grammateus: from Ancient Greek Papyri to a Web Application

Elisa Nury[1]

1 Université de Genève, Switzerland

*Corresponding author: Elisa Nury elisa.nury@unige.ch

Abstract
This paper describes the workflow of the Grammateus project, from gathering data on Greek documentary papyri to the creation of a web application. The first stage is the selection of a corpus and the choice of metadata to record: papyrology specialists gather data from printed editions, existing online resources and digital facsimiles. In the next step, this data is transformed into the EpiDoc standard of XML TEI encoding, to facilitate its reuse by others, and processed for HTML display. We also reuse existing text transcriptions available on <http://papyri.info/>. Since these transcriptions may be regularly updated by the scholarly community, we aim to access them dynamically. Although the transcriptions follow the EpiDoc guidelines, the wide diversity of the papyri as well as small inconsistencies in encoding make data reuse challenging. Currently, our data is available on an institutional GitLab repository, and we will archive our final dataset according to the FAIR principles.

keywords
digital humanities; classics; tei; exist-db; papyrus; typology

INTRODUCTION

Documentary papyri record everyday life, both public and private, in Ptolemaic, Roman and Byzantine Egypt. Private documents include for instance contracts, accounts, letters; whereas public documents are related to government administration or to the relation between government and individuals who petition an official for a request to be granted [Palme 2009].

The Grammateus project aims at providing a new typology of documentary papyri, taking into account not only the text, but also material aspects of a document such as its dimensions, its shape, and how the text is laid out on the page. In summary, how did a trained scribe prepare a document? Instead of approaching the problem from a modern point of view, we are trying to take the scribe's perspective.

To achieve this goal, there are several important digital papyrological resources available online, encoded according to the TEI [EpiDoc Guidelines], that we can reuse. The most important ones are the *Heidelberger Gesamtverzeichnis der Griechischen Papyrusurkunden Ägyptens* [HGV] and the *Duke Databank of Documentary Papyri* which are both accessed through [papyri.info], thanks to unique identifiers.

This contribution describes in more detail how we select and prepare our data, and how we process it. The workflow is iterative: we began with a very limited corpus, and have started to expand it progressively. For this reason, we must take great care that every stage of the workflow is well integrated and can be repeated easily each time we add new papyri to our corpus.

I DATA SELECTION

The first part of our workflow consists of selecting the corpus of papyri to enter into our database, as well as which metadata to record. This is mostly the responsibility of the papyrologists in the team, although we already need to think of the most suitable digital format in which to record our data for further processing.

1.1 The Corpus

Documentary papyri represent a corpus of about sixty thousand published documents, coming mostly from Egypt between 400 BCE and 650 CE. Our purpose is not to be exhaustive, but rather to study a representative sample of each type of document. Therefore, we select documents from Graeco-Roman Egypt (300 BCE – 300 CE). In addition, a facsimile must be available either online or in a printed publication, so that papyrologists may be able to see the overall aspect of the papyrus as well as how the text is laid out. Finally, papyri should be complete documents, in order to obtain precise dimensions which cannot be extrapolated from fragments.

1.2 The Metadata

For each papyrus, a set of metadata is recorded by a papyrologist into a matrix where each row represents one papyrus. Although the papyrologist works with a Microsoft Excel spreadsheet for convenience, we save the matrices in tab-separated text files, to make versioning easier.

The metadata include identifiers and canonical references - essential for accessing other databases, dimensions, direction of fibres, image links, and information pertaining to the text (such as margins, blank spaces or text sections). These metadata are gathered from printed publications, online resources and digital facsimiles. A more detailed description of our data can be found on the project repository.[1]

1.3 In Practice

HGV EpiDoc data[2] provides a set of descriptive keywords in XML which allow us to filter for specific types of document. These keywords are not always consistent and may be in several languages. From a local copy of the HGV data, we can query the XML for files that contains certain keywords:

```
for $doc in collection("/db/apps /HGV/")
where ($doc//tei:term[contains(text(), 'List')] or $doc//tei:term[contains(text(),
'Liste')])
```

This lets us find most lists from the papyri available in HGV. However, there are close to five thousand lists which need to be examined, and we already know that a large number will be fragments and therefore will not fit our criteria. The fact that a document is fragmentary is not encoded in HGV, but can be inferred fairly reliably from the transcription: the presence of a gap of unknown length means that the text is likely fragmentary. Therefore, we filter the lists again to remove those for which the corresponding transcription contains a <gap> element with @reason ‘lost’ and @extent ‘unknown’. We also remove the documents that have neither a link

[1] <https://gitlab.unige.ch/papyri-dev/grammateus/blob/master/ > (Accessed Feb. 2020).

[2] <https://github.com/papyri/idp.data/tree/master/HGV_meta_EpiDoc> (Accessed Aug. 2019).

to a digital facsimile or a reference to a printed image. This lowers the number of lists to roughly 800, which reduces the time-consuming task of data selection and preparation.

Once the papyri have been selected and the metadata entered into the spreadsheet, a Python script is used to transform each row into an EpiDoc file and saved into the data repository of the project.

II DATABASE AND INFRASTRUCTURE

The next step is to create a web application to visualise our data. We use [eXist], a platform that is designed for XML Databases. With XSLT and XQuery technologies, the data can be searched and displayed in HTML format, and the external data can be accessed.

The data are organised as follow: in a local repository the matrices, the EpiDoc files, and the eXist application are stored, and regularly synchronized with a gitlab repository on the university's servers. The eXist instance is also synchronized with the local repository in both directions: a XQuery script loads the XML data files into the database, and the app is regularly exported to the local repository.

As transcriptions and other metadata may be updated occasionally on [papyri.info], these external records should be accessed dynamically in order to avoid outdated information on the Grammateus web app. However, it is not possible to index external files, which in turns makes it impossible to search for date and place for instance, which are essential to papyrological research. Therefore, we make a copy of the <origin> TEI element from [HGV], and update it once a month. It is possible to perform this task automatically thanks to eXist's scheduler facility.

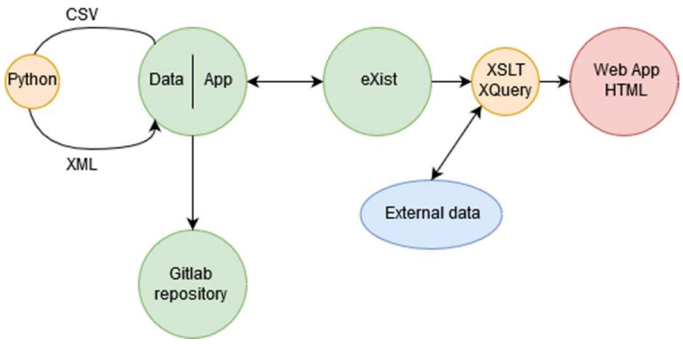

Figure 1. A summary of the current infrastructure.

III DISPLAYING PAPYRI IN HTML

For each papyrus, we display an HTML model consisting of three juxtaposed layers:

- A brown rectangle in the background representing the papyrus, with stripes indicating the direction of fibres.
- The textual transcription extracted from [papyri.info] and transformed in HTML with the [EpiDoc Stylesheets].
- Text sections highlighted in colors, line by line.

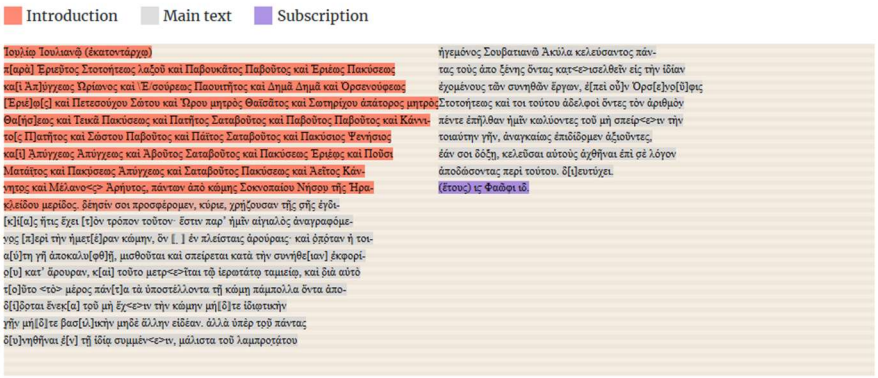

Fig. 2. The HTML representation of p.gen. 2.1 16.

It might seem that displaying the text in the same order as it is written on papyrus, and counting lines of transcription would be easy enough tasks. However, the large variety of papyri, as well as the many possible encodings, can make the reuse of transcription challenging. For instance, we may have a good image while the transcription is actually that of a copy, where linebreaks do not correspond. It may also happen that the transcription is not a faithful reproduction of the text on the papyrus.[3]

Counting lines for the text sections is not a matter of counting TEI <lb> or HTML
 elements, and one <lb> in TEI does not necessarily equals one
 in the HTML display.[4] Lines are not always consistently numbered: they may be continuously numbered from the first to the last column, but it can happen that the numbering starts again at 1 in the next column (for instance papyrus BGU 3 996).

Because of these issues, we have modified the [EpiDoc Stylesheets] so that lines in the final HTML display are always continuous and match as closely as possible the display on [papyri.info].

Conclusion

The Grammateus project is exploring the typology of Greek documentary papyri with a new approach, and therefore the corpus is constantly evolving. Our workflow takes this into account, by making the whole process reproducible, and by preparing for the regular addition of new documents into the the database. For instance, we have added a group of 79 papyri in a matter of minutes, thanks to the combination of Python and XQuery scripts. If the EpiDoc format had to be changed, or new metadata added, those scripts would be easily adapted to update the whole database.

References

EpiDoc Guidelines https://www.stoa.org/epidoc/gl/latest/
EpiDoc Stylesheets https://sourceforge.net/p/epidoc/wiki/Stylesheets/
eXist http://exist-db.org
HGV http://aquila.zaw.uni-heidelberg.de
Palme, B. The Range of Documentary Texts: Types and Categories. *The Oxford Handbook of Papyrology*, ed. Bagnall, R S. Oxford University Press (Oxford), 2009: 358–94. DOI: 10.1093/oxfordhb/9780199843695.013.0016.
Papyri.info http://papyri.info/

[3] See for instance P.Dion. 16, where the lines appearing on top of the papyrus are transcribed at the end of the text. In other cases, it may be that the text is copied twice on the papyrus, but transcribed only once.

[4] The transformation to HTML is affected by textual phenomena such as *paragraphos*, *vacat*, or seals.