Greek Spaces in Roman Times: the Construction of Greek Geography in Pliny's Naturalis Historia

A Data Management Plan created using DMPonline.be

Creator: Laura Soffiantini

Affiliation: KU Leuven (KUL)

Funder: Fonds voor Wetenschappelijk Onderzoek - Research Foundation Flanders (FWO)

Template: FWO DMP (Flemish Standard DMP)

Principal Investigator: Laura Soffiantini

Data Manager: Laura Soffiantini

Project Administrator: Laura Soffiantini

Grant number / **URL:** 1193123N

ID: 197053

Start date: 01-11-2022

End date: 31-10-2024

Project abstract:

The influence of Greek culture on Pliny's the Elder *Naturalis Historia* has been largely recognized over the last few years. However, an overall analysis of Pliny's geographical discourse on Greece is still lacking. The aim of the project is to provide a comprehensive insight into Pliny's description of Greece. Combining the criteria of the history of knowledge inspired by Östling et al. (2018) with the Digital Classical Geography of Palladino (2021), I will reconstruct the spatial model of Greece to understand how Pliny's perspective influenced its overall image. The investigative strategy will be based on an innovative twofold methodology. First, I will employ sentiment analysis and discourse analysis to investigate the linguistic and narrative devices that Pliny uses to characterize Greece. Second, I will examine the spatial connectivity between Greek spaces and the relational structure according to which Pliny disseminates knowledge about them throughout *NH*. This will result in a database of Greek geographical entities which will be encoded in a multi-layered digital map of the Greek spatial model in the *NH*–a tool that will support the visualization and investigation of connections running across Greek spaces. Finally, the analysis of the position of Greece in Pliny's geographical and political discourse will contribute to shed new light on the Roman perspective on Greece during the Early Imperial history and during the Flavian age more specifically.

Last modified: 18-04-2023

Greek Spaces in Roman Times: the Construction of Greek Geography in Pliny’s Naturalis Historia
Application DMP

Questionnaire

Describe the datatypes (surveys, sequences, manuscripts, objects …) the research will collect and/or generate and /or (re)use. (use up to 700 characters)

The research will consist both in reusing and generating data (exclusively not personal data).
Input sources:

1. digitalized versions of the text (.html). Multiple digitalized versions of Pliny's text are currently available online both in Latin and in translation (English). In particular, in the first steps of my project I will largely rely upon an annotated version of the text from ToposText.org, a project which aims at mapping places in the ancient literature. My scope is to reuse the annotations of places in ToposText (tags in the html file) as well as additional information such as geographic coordinates, links to other dataset, unique IDs nested in ToposText tags.
2. Print editions and indexes (paper). Print editions of the text will be a reference point, since some of the information contained there (i.e., critical apparatus) are not available in the digitalized versions of Pliny's text. In addition, one of the steps of my project will consist in the transformation of the indexes (i.e., lists of things, places, and persons mentioned in the text) from paper to pdf via OCR. In this way, I will be able to in order to process and reuse indexes as authoritative lists during the extraction process.
3. bibliography (.pdf, paper).
4. Latin sentiment lexicon. In order to perform sentiment analysis to the text, I will rely upon an existing sentiment lexicon in Latin, that is a dictionary of words with a score (positive, neutral, negative). Reusing the evaluation scores assigned to Latin words will permit me to analyze the general sentiment associated to places in the text.
5. Ontologies to describe relationships between places. In order to describe the relations between spaces in a standardized and machine-actionable way, I will translate spatial relations into logic statements by using and developing ontologies created to analyze spatial concepts.

Outputs:

1. Python scripts (.ipynb). One of the most consistent outputs of my project will be the creation of code scripts to perform specific tasks such as performing Named Entity Recognition to extract place names, performing Network Analysis on co-occurring words, performing sentiment analysis and topics modelling. In comparison to existing scripts on these topics, I will focus on a specific subject (space) in an ancient literary text written in Latin.
2. CSV files. The output data will usually be organized in a tabular format which will permit to deal with a large amount of data. The CSV files include, for instance, the list of all place names in Pliny's encyclopedia, the list of Latin passages containing place names, and the list of co-occurring place names to perform Network Analysis.
3. A FileMaker relational database. In order to manage the large amount of CSV files produced, I will generate a FileMaker database which will permit me to explore in a more practical way all the different data collected about space.
4. A network (NetworkX). One of the outputs of my research will be the creation of a network of co-occurrences that will permit to explore which places are connected to each other and how. My aim is to create an interactive network which will allow to explore Pliny's massive text and to investigate the interconnections in the encyclopedia.
5. A multi-layered map (NodeGoat). Since the project has a focus on places, I will generate a map that will permit to visualize the places mentioned by Pliny. The map will also permit to perform further analysis, for instance investigating which areas are described in more detail.

My data will probably be less than 50GB in total.

Specify in which way the following provisions are in place in order to preserve the data during and at least 5 years after the end of the research? Motivate your answer. (use up to 700 characters)

During my research project, I will have access to the OneDrive storage service where the research data will be archived guaranteeing the storage capacity required for my data. All the relevant research data will be also stored on Zenodo (https://zenodo.org/) in an open access policy and without copyright-restrictions. This will permit me to deposit in a safe and secure repository all the tabular information and the datasets I will generate during my research (including data for the verification of research results) assigning a persistent identifier to them. Choosing Zenodo, an established and trusted resource, will satisfy the requirements in terms of longevity and file preservation of at least 10 years. As for the code scripts (data with potential re-use), all the relevant research data will be stored in a GitHub profile (https://github.com/). This will permit to share my project in the largest public source code host. Moreover, by connecting my GitHub account with Zenodo new versions of the code will be automatically archived in Zenodo. In this way, the code will get a DOI to make it more citeable. I am also in contact with the team working at the ToposText project (https://topostext.org/), whose data and metadata will be reused in my project during the first steps of my workflow. The feedbacks on the ToposText data (ie., corrections on the tag annotations) will be integrated and preserved in their website. Moreover, the spatial information that I will collect (ie, place names) will be integrated in the Trismegistos database and, more specifically, in the discipline-specific Trismegistos Geo repository (https://www.trismegistos.org/geo/. This will permit to enrich the Trismegistos database and to preserve my research data with the guarantee of a sufficient storage capacity. Trismegistos will also host the network I will generate in Trismegistos Tomatillo, an online environment which permit to visualize and customize network visualization.

What’s the reason why you wish to deviate from the principle of preservation of data and of the minimum preservation term of 5 years? (max. 700 characters)

I dont' wish to deviate from the minimum preservation term of 10 years.

Are there issues concerning research data indicated in the ethics questionnaire of this application form? Which specific security measures do those data require? (use up to 700 characters)

There are no issues concerning my research data.

Which other issues related to the data management are relevant to mention? (use up to 700 characters)

There are no other issues related to the data management.

Greek Spaces in Roman Times: the Construction of Greek Geography in Pliny’s Naturalis Historia
DPIA

DPIA

Have you performed a DPIA for the personal data processing activities for this project?

- Not applicable

Greek Spaces in Roman Times: the Construction of Greek Geography in Pliny’s Naturalis Historia
GDPR

GDPR

Have you registered personal data processing activities for this project?

- Not applicable

Greek Spaces in Roman Times: the Construction of Greek Geography in Pliny's Naturalis Historia

FWO DMP (Flemish Standard DMP)

1. Research Data Summary

List and describe all datasets or research materials that you plan to generate/collect or reuse during your research project. For each dataset or data type (observational, experimental etc.), provide a short name & description (sufficient for yourself to know what data it is about), indicate whether the data are newly generated/collected or reused, digital or physical, also indicate the type of the data (the kind of content), its technical format (file extension), and an estimate of the upper limit of the volume of the data.

				Only for digital data	Only for digital data	Only for digital data	Only for physical data
Dataset Name	Description	New or reused	Digital or Physical	Digital Data Type	Digital Data format	Digital data volume (MB/GB/TB)	Physical volume
		Please choose from the following options: • Generate new data • Reuse existing data	*Please choose from the following options:* • Digital • Physical	*Please choose from the following options:* • Observational • Experimental • Compiled/aggregated data • Simulation data • Software • Other • NA	*Please choose from the following options:* • .por, .xml, .tab, .cvs,.pdf, .txt, .rtf, .dwg, .gml, … • NA	*Please choose from the following options:* • <100MB • <1GB • <100GB • <1TB • <5TB • <10TB • <50TB • >50TB • NA	
Digital bibliography	Books, articles, papers concerning Pliny's work and the topic of geography and spatial representation in Antiquity	Reuse existing data	Digital	Compiled data	.pdf	<1 GB	
Print bibliography	Books, articles, papers concerning Pliny's work and the topic of geography and spatial representations in Antiquity	Reuse existing data	Physical	Compiled data	NA	<1 GB	Archived in KU Library
Print edition	Critical editions of Pliny's text	Reuse existing data	Physical	Other	NA	<1 GB	Archived in KU Library
Digitalized Latin text	Digitalized version of Pliny's text	Reuse existing data	Digital	Other	.html	<1 GB	
Digitalized English text	English annotated translation of Pliny's text	Reuse existing data	Digital	Compiled data	.html	< 1 GB	
Place coordinates	Geographic coordinates of ancient places from existing repositories (Trismegistos, GeoNames)	Reuse existing data	Digital	Compiled data	.csv	< 1 GB	
Lemmatized Latin version	Lemmatized version of Pliny's text	Reuse existing data	Digital	Compiled data	.xml	< 1 GB	
Indexes	Indexed of Pliny's books (from paper to txt)	Reuse existing data	Digital	Compiled data	.pdf, .txt	< 1 GB	
Ontology	Descriptive terminology to describe relations between spaces	Reuse existing data	Digital	Compiled data	.csv	< 1 GB	
Latin sentiment lexicon	Lexicon associating a score (positive, neutral, negative) to each Latin word	Reuse existing data	Digital	Compiled data	.csv	< 1 GB	
Research journal	Daily updates on the steps of the research	Generate new data	Digital	Compiled data	.doc	< 1 GB	
Notes	Notes on the bibliography and readings	Generate new data	Digital	Compiled data	.doc	< 1 GB	
Python script - extraction of place names	The script permits to extract place names, position (book, chapter), external IDs and additional information from the html source page	Generate new data	Digital	Other	.ipynb	< 1 GB	
Python script - Named Entity Recognition (NER)	The script permits to perform NER on the digitalized version in translation and evaluate the results of the extraction	Generate new data	Digital	Other	.ipynb	<1 GB	
Python script - correspondence between Latin and English texts	The script permits to associate the English and Latin version of the digitalized texts	Generate new data	Digital	Other	.ipynb	< 1 GB	
Python script - co-occurrences of place names	The script permits to analyze place names co-occurring in the same text window (ie. same paragraph)	Generate new data	Digital	Other	.ipynb	< 1 GB	
CSV file - dataset of place names	Dataset of place names containing the place name, coordinates, position (book, chapter, paragraph), coordinates (lat, long), unique IDs (ToposText, Trismegistos, Wikidata)	Generate new data	Digital	Compiled data	.csv	< 1 GB	
CSV file - dataset of textual sections	The dataset includes the Latin sentences containing a reference to a place name	Generate new data	Digital	Compiled data	.csv	< 1 GB	
CSV file - evaluation of text polarity	Nouns and adjectives associated to place names are scored by the Latin sentiment lexicon	Generate new data	Digital	Compiled data	.csv	< 1 GB	
Map	The multi-layered map displays the places mentioned in Pliny's text	Generate new data	Digital	Other		< 100 GB	
Network	The network measures how the places are interconnected to each other on the basis of their co-occurrences in the same text window and describes which places are more frequently linked to other places	Generate new data	Digital	Other		< 100 GB	
Model	Descriptive model of Pliny's geographic and spatial description of Greece	Generate new data	Digital	Other		< 100 GB	

If you reuse existing data, please specify the source, preferably by using a persistent identifier (e.g. DOI, Handle, URL etc.) per dataset or data type:

- Digitalized annotated English translation: ToposText (https://topostext.org/work/148)
- Lemmatized Latin version: LascivaRoma (https://github.com/lascivaroma/latin-lemmatized-texts/blob/main/lemmatized/xml/urn:cts:latinLit:phi0978.phi001.perseus-lat2.xml)
- Place coordinates: Trismegistos Place (https://www.trismegistos.org/geo/)
- Ontology: GeoLat (https://geolat.uniupo.it/the-project)
- Latin sentiment lexicon: https://zenodo.org/record/3862149#.Y-uzli1aZQI

Are there any ethical issues concerning the creation and/or use of the data (e.g. experiments on humans or animals, dual use)? Describe these issues in the comment section. Please refer to specific datasets or data types when appropriate.

- No

Will you process personal data? If so, briefly describe the kind of personal data you will use in the comment section. Please refer to specific datasets or data types when appropriate.

- No

Does your work have potential for commercial valorization (e.g. tech transfer, for example spin-offs, commercial exploitation, …)? If so, please comment per dataset or data type where appropriate.

- No

Do existing 3rd party agreements restrict exploitation or dissemination of the data you (re)use (e.g. Material/Data transfer agreements/ research collaboration agreements)? If so, please explain in the comment section to what data they relate and what restrictions are in place.

- No

Are there any other legal issues, such as intellectual property rights and ownership, to be managed related to the data you (re)use? If so, please explain in the comment section to what data they relate and which restrictions will be asserted.

- No

2. Documentation and Metadata

Clearly describe what approach will be followed to capture the accompanying information necessary to keep data understandable and usable, for yourself and others, now and in the future (e.g., in terms of documentation levels and types required, procedures used, Electronic Lab Notebooks, README.txt files, Codebook.tsv etc. where this information is recorded).

Documentation will be particularly relevant for the datasets generated during my research and for the code scripts. As for the code scripts, the information necessary to keep data understandable and usable will be recorded in the Jupyter Notebook in the form of code annotations in order to describe the procedural information, definition of variables, packages and libraries used during the analysis. The documentation will also be described in my thesis and in the form of README.txt files in the the repository (GitHub) where the scripts will be stored and shared. As for the datasets, documentation will be recorded during every step of my research in my research journal, in my thesis and also in the repository (Zenodo) where datasets generated will be stored and shared.

Will a metadata standard be used to make it easier to find and reuse the data? If so, please specify (where appropriate per dataset or data type) which metadata standard will be used. If not, please specify (where appropriate per dataset or data type) which metadata will be created to make the data easier to find and reuse.

- Yes

Metadata will be used to describe the collected data. In the dataset of place name I will use unique IDs for ancient place names already in use in existing repositories (ToposText, GeoNames, Trismegistos) to identify and disambiguate places and make the data easier to find. In addition, text division (book, chapter, paragraph) will be imported from the critical edition of Pliny's text which will permit to navigate from the Latin version and the English translation of the text. Attestations of place names will be also assigned a unique ID, that is a stable ID to identify the attestation of a place in a particular section of the work. Finally, each token in the lemmatized version will be associated with a unique ID which will permit to find and disambiguate each token.

3. Data storage & back-up during the research project

Where will the data be stored?

During my entire research project, I will have access to the OneDrive for Business could storage service via KU Leuven Microsoft 365 where the research data will be archived guaranteeing the storage capacity required for my data.

How will the data be backed up?

I use standard back-up provided by KU Leuven ICTS.

Is there currently sufficient storage & backup capacity during the project? If yes, specify concisely.
If no or insufficient storage or backup capacities are available, then explain how this will be taken care of.

- Yes

The maximum storage and backup capacity size of the OneDrive for Business of 2 TB is largely sufficient for the estimated value of my research data.

How will you ensure that the data are securely stored and not accessed or modified by unauthorized persons?

The security level of the OneDrive for Business could storage is evaluated as medium in the KU Leuven interactive storage guide. In addition, multifactor authentication with the KU Leuven Authenticator app is activated ensuring that data are not accessed or modified by unauthorized persons.

What are the expected costs for data storage and backup during the research project? How will these costs be covered?

OneDrive for Business is free for staff and students of KU Leuven.

4. Data preservation after the end of the research project

Which data will be retained for at least five years (or longer, in agreement with other retention policies that are applicable) after the end of the project? In case some data cannot be preserved, clearly state the reasons for this (e.g. legal or contractual restrictions, storage/budget issues, institutional policies...).

Datasets, code scripts, network, map generated during my research will be retained for at least ten years after the end of the project.

Where will these data be archived (stored and curated for the long-term)?

Tabular research data and datasets generated during my research project will be stored on Zenodo assigning a persistent identifier to them. The code scripts will be stored in a GitHub profile (https://github.com/). Feedback and corrections on the existing ToposText project (on which I will largely rely during the first steps of my research) will be integrated and preserved in ToposText website. Finally, Trismegistos will host the network I will generate in Trismegistos Tomatillo, an online environment which permit to visualize and customize network visualization.

What are the expected costs for data preservation during the expected retention period? How will these costs be covered?

No costs are expected at the moment for data preservation.

5. Data sharing and reuse

Will the data (or part of the data) be made available for reuse after/during the project? In the comment section please explain per dataset or data type which data will be made available.

- Yes, in an Open Access repository

Code scripts: all the relevant Python code scripts generated during the project will be made available with documentation.
Tabular data generated during the project (ie., dataset of place names, dataset of textual passages) will be made available.

If access is restricted, please specify who will be able to access the data and under what conditions.

The access is not restricted.

Are there any factors that restrict or prevent the sharing of (some of) the data (e.g. as defined in an agreement with a 3rd party, legal restrictions)? Please explain in the comment section per dataset or data type where appropriate.

- No

Where will the data be made available? If already known, please provide a repository per dataset or data type.

Code scripts: GitHub.
Datasets: Zenodo.

When will the data be made available?

Some of the data will be made available during the project. Other data will be made available upon publication of research results.

Which data usage licenses are you going to provide? If none, please explain why.

I am going to provide OpenAccess usage licence.

Do you intend to add a PID/DOI/accession number to your dataset(s)? If already available, you have the option to provide it in the comment section.

- Yes

What are the expected costs for data sharing? How will these costs be covered?

The sevices used for data sharing (Zenodo, GitHub, Tomatillo) are free of charge.

6. Responsibilities

Who will manage data documentation and metadata during the research project?

Me (PhD candidate)

Who will manage data storage and backup during the research project?

Me (PhD candidate)

Who will manage data preservation and sharing?

Me (PhD candidate)

Who will update and implement this DMP?

Me (PhD candidate), my supervisor (prof. Margherita Fantoli).