NIKAW (Networks of Ideas and Knowledge in the Ancient World)

A Data Management Plan created using DMPonline.be

Creators: Margherita Fantoli, Jens Bürger

Affiliation: KU Leuven (KUL)

Template: KU Leuven BOF-IOF

Principal Investigator: Margherita Fantoli

Project Administrator: Jens Bürger

Grant number / **URL:** IDN/22/001

ID: 197883

Start date: 01-03-2023

End date: 28-02-2027

Project abstract:

The ID-N project NIKAW (Networks of Ideas and Knowledge in the Ancient World) aims at bringing together the fields of Humanities, Social Sciences and AI. We will analyze the circulation of knowledge in the ancient world through the study of the mention of names (both of authors and other historical figures) in the corpus of Latin and Ancient Greek literary texts. Using automated NLP methodologies, we will reconstruct the network of mentions in ancient texts and, building on the methodologies of economics and social network analysis, we will analyze the flow of intellectual influences through the network, and the impact of well-known historical events on this flow. We aim at building a methodological framework for applying models developed for present-day networks on data of antiquity; improving the combination of different AI methods; and providing new insights on the knowledge circulation in the ancient world, with a special focus on the advent of Christianity.

Last modified: 31-03-2023

NIKAW (Networks of Ideas and Knowledge in the Ancient World)

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.

Dataset name / ID	Description	New or reuse	Digital or Physical data	Data Type	File format	Data volume	Physical volume
		Indicate: ***N****(ew data) or* ***E****(xisting data)*	Indicate: **D**(igital) or **P**(hysical)	Indicate: **A**udiovisual **I**mages **S**ound **N**umerical **T**extual **M**odel **SO**ftware Other (specify)		Indicate: <1GB <100GB <1TB <5TB >5TB NA	
Trismegistos	Metadata database of the ancient world	E	D	T		<100GB	
Topostext	Indexed collection of ancient texts and mapped places relevant the the history and mythology of the ancient Greeks	E	D	T			
LASLA	130 Classical Latin texts	E	D	T		<1GB	
LGPN	Lexicon of Greek Personal Names	E	D	T		<1GB	
EDCS	Epigraphic Database	E	D	T		<1GB	
GLAUx	Greek language corpus	E	D	T			
Latin Lemmatized Texts	Latin language corpus	E	D	T			
Digital Athenaeus	Catalog of Authors and Works cited in the text of the *Deipnosophists*	E	D	T		<1GB	
Large set of name mentions	Mentions extracted from the listed databases via the language models developed within NIKAW	N	D	T		tbd	
Enriched annotation of the Greek and Latin corpus	Additional annotations for the recognition of persons referred to by multiple names or expressions	N	D	T		<1GB	
VIAF	Name authority files into a single OCLC-hosted name authority service	E	D	T			
Prosopographia Imperii Romani	Person encyclopedia of the leading class of the Roman Empire in the early and high imperial period	E	D	T			
Graph database	Graph database integrating relevant data from existing databases with data newly extracted/generated by the NIKAW project.	N	D	T	neo4j	tbd	
Disambiguated entities	A database of unique people with links to their mentions	N	D	T		tbd	
Citation network	Graph database of people and citations demonstrating intellectual influences	N	D	T	neo4j	tbd	

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:

Dataset name / ID	Persistent identifier
Trismegistos	https://www.trismegistos.org/
Topostext	https://topostext.org/
LASLA	https://www.lasla.uliege.be/cms/c_8508894/fr/lasla
LGPN	https://www.lgpn.ox.ac.uk/
EDCS	https://db.edcs.eu/epigr/hinweise/hinweis-en.html
GLAUx	https://aclanthology.org/2021.lchange-1.6/
Latin Lemmatized Texts	https://github.com/lascivaroma/latin-lemmatized-texts
Digital Athenaeus	https://www.digitalathenaeus.org/
VIAF	https://viaf.org/
Prosopographia Imperii Romani	https://pir.bbaw.de/#/overview

Are there any ethical issues concerning the creation and/or use of the data (e.g. experiments on humans or animals, dual use)? If so, refer to specific datasets or data types when appropriate and provide the relevant ethical approval number.

- No

Will you process personal data? If so, please refer to specific datasets or data types when appropriate and provide the KU Leuven or UZ Leuven privacy register number (G or S number).

- 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 or 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.

- Yes

Some annotated corpora require Data transfer Agreements at the moment (e.g. LASLA). They will be signed if necessary when the corpora are used.

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

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).

All newly generated data will be stored in file formats common for the respective data types and research communities. The data will be documented in accordance to data type conventions and standards (e.g., graph data models, UML schema). All relevant information will be described in corresponding readme files.

Large set of name mentions	This data will be collected from the existing data sources via Named Entity Recognition.
Enriched annotation of the Greek and Latin corpus	The new annotations are derived from insights gained from the interplay between the bi-lingual language models and the graph database.
Graph database	follow Neo4j standards for documentation
Disambiguated entities	We will provide clear indications of the entities IDs and how they are linked to textual IDs
Citation network	follow Neo4j standards for documentation

Will a metadata standard be used to make it easier to find and reuse the data?
If so, please specify which metadata standard will be used.

If not, please specify which metadata will be created to make the data easier to find and reuse.

- Yes

When applicable, we will implement standard metadata, such as TEI-XML for textual data or Dublin Core.

Data Storage & Back-up during the Research Project

Where will the data be stored?

- OneDrive (KU Leuven)
- Other (specify below)
- Sharepoint online

The hosting and storage of the graph database (server + data) is still to be defined. However, a backup of the data will also be stored via OneDrive or Sharepoint.

How will the data be backed up?

- Standard back-up provided by KU Leuven ICTS for my storage solution

Is there currently sufficient storage & backup capacity during the project?

If no or insufficient storage or backup capacities are available, explain how this will be taken care of.

- Yes

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

Data will be stored on KU Leuven infrastructure. Any data stored outside of KU Leuven (potentially the graph database) will be secured via access control.

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

The graph database will be hosted on a virtual server for which 8000 euros are budgeted.

Data Preservation after the end of the Research Project

Which data will be retained for 10 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...).

- All data will be preserved for 10 years according to KU Leuven RDM policy

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

- KU Leuven RDR
- Other (specify below)

Zenodo

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

Standard KU Leuven services + free of use repositories (Zenodo) will not require additional funding.

Data Sharing and Reuse

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

- Yes, as open data

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

NA

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 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.

- KU Leuven RDR (Research Data Repository)

When will the data be made available?

- Upon publication of research results

Which data usage licenses are you going to provide?

If none, please explain why.

- CC-BY 4.0 (data)

Do you intend to add a persistent identifier (PID) to your dataset(s), e.g. a DOI or accession number? If already available, please provide it here.

- Yes, a PID will be added upon deposit in a data repository

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

No costs involved (RDR used)

Responsibilities

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

The coordinating team (Jens Bürger and Margherita Fantoli) and the two PhD scholars (Evelien De Graaf and Marijke Beersmans)

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

The PIs supervising the PhDs generating the data, and the PhDs themselves.

Who will manage data preservation and sharing?

Jens Bürger and Margherita Fantoli

Who will update and implement this DMP?

Jens Bürger and Margherita Fantoli