

MECANO: The Mechanics of Canon Formation and the Transmission of Knowledge from Greco-Roman Antiquity

Data Management Plan

22/06/25

v.1.1: first version of DMP for all 10 subprojects

Table of contents

1.	Citations and quotations in the Naturalis Historia: creating the canon in the Encyclopaedia.....	- 2 -
2.	The philosophical canon and the art of (mis)quoting Plato and Aristotle in the CAG.....	- 9 -
3.	The presence of classics in early modern book history - Data management plan (after 6 months).....	- 19 -
4.	Pulse and Physiology in Hellenistic Science.....	- 22 -
5.	Detecting and Retrieving Lost Historians.....	- 28 -
6.	Recovering anonymous late-antique preachers in the corpus of pseudo-Augustinian sermons: Data Management Plan.....	- 34 -
7.	Syntax, formulaic structures, and canon-marking in Greek and Arabic: documentary texts and Galen -	40 -
8.	Ancient sources on matter in Late Medieval Commentaries on Aristotle.....	- 46 -
9.	Contextual scientometrics—Uncovering and understanding referencing patterns to the ancient canon in modern scholarly discourses.....	- 58 -
10.	A democratic turn? Uncovering and understanding references to Graeco-Roman antiquity in 20th-century French public discourse.....	- 65 -

1. Citations and quotations in the *Naturalis Historia*: creating the canon in the Encyclopaedia

Full DMP

Version information

Action number

101120349

Action acronym

MECANO

Action title

MECANO: The Mechanics of Canon Formation and the Transmission of Knowledge from Greco-Roman Antiquity

DMP version number

v1.1

Date

17/02/2025

1. Data summary

1.1 Will you re-use any existing data and what will you re-use it for?

In my project I will re-use the following data:

Digital corpus of Pliny the Elder *Naturalis Historia*, stored in a collaborative relational database on FileMaker. The text was OCR-generated, and then automatically annotated with regard to lemmatization and morphology. I will re-use the text as a starting point for my own work.

- Physical and digital books and articles.
- Jupyter notebooks. I will re-use Python Jupyter Notebooks written by colleagues or by myself, to work on the digital text.

1.2 What types and formats of data and other research outputs will the project generate or re-use?

- Source data: Digital corpus of Pliny the Elder *Naturalis Historia*, stored in a relational database on FileMaker (cf. section 1.1). The data is in .fmp12 format, but can be exported in .csv, or .tab.

- Output data: Final annotated corpus of Pliny the Elder *Naturalis Historia*, 2-6. The corpus will be fully annotated with regard to Named Entity Recognition. The data will be available both in .fmp12 format, and .csv, or .tab. The dataset will be published.
- Source data: Physical and digital books and articles.
- Source and output data: Jupyter notebooks. .ipynb format.

1.3 What is the purpose of the data generation or re-use and its relation to the objectives of the project?

The main research question is to investigate the correlation between the citation of people and the textual and linguistic features in the books 2-6 of the *Naturalis Historia*. To answer to this question, it will be first necessary to implement a digital annotation of the involved books (stored in the dataset mentioned in section 1.1 and 1.2). This annotation will regard in particular the lemmas and the part of speech of the text and the Named Entities (that is, in this case, the people cited by the author). The books already feature an automatic morphological annotation, that need to be checked and corrected whenever necessary. The Named Entity annotation will be performed automatically and then manually checked. The annotated corpus will be finally linked to the LiLa Knowledge Base (<https://lila-erc.eu/>) and/or to the Pauly-Wissowa's *Realencyclopädie der classischen Altertumswissenschaft* and/or to other knowledge bases (such as Wikidata). After the annotation is performed, it will be possible to check how the citations and quotations of people are distributed in the books, and if there are any linguistic patterns that emerge when Pliny cites them. The linking to LiLa will be possibly used for finding any other relevant feature of the citations, while the linking to the other knowledge bases will allow to disambiguate the people cited. The final objective will be to outline the structure of the canon of the *Naturalis Historia* astronomical and geographical books and to delineate the mechanisms that guided its creation.

1.4 What is the expected size of the data that you intend to generate or re-use?

- Source data: Digital corpus of Pliny the Elder *Naturalis Historia*, stored in relational database on FileMaker. The size is around 50 MB.
- Output data: Final annotated corpus of Pliny the Elder *Naturalis Historia*, 2-6.

The total size of all the data should not be more than 50 GB.

1.5 What is the origin/provenance of the data, either generated or re-used?

- Source data: Digital corpus of Pliny the Elder *Naturalis Historia*. <https://doi.org/10.5281/zenodo.4337145>. <https://github.com/lascivaroma/latin-lemmatized-texts>. The dataset is open access
- Output data: Final annotated corpus of Pliny the Elder *Naturalis Historia*, 2-6. The already provided annotation will be first checked on FileMaker. NER will be performed automatically with LatinBERT and LatinCY (and possibly other models).

1.6 To whom might your data be useful ('data utility'), outside your project?

The dataset will be useful for other Latin and Digital Humanities scholars that want to exploit the annotation provided for their research questions, or want to use it as a starting point for other NLP tasks.

2.1 FAIR data: Making data findable, including provisions for metadata

2.1.1 Will data and other research outputs be identified by a persistent identifier?

- Yes: describe below

Yes, the dataset will be published with a DOI on Zenodo, and the author will be identified by an ORCID id.

The dataset will also be identified by a CTS URN and linked to the LiLa Knowledge base, where each item has its own URI. The linking to external knowledge bases will provide an id for each person cited.

**2.1.2 Will rich metadata be provided to allow discovery?
What metadata will be created?
What disciplinary or general standards will be followed?
In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how.**

Data citation metadata will be added when the dataset will be published (title, description, date...).
A .README file will contain all the relevant information.

CTS URNs standard will be used to reference the text of the dataset.

Trismegistos ids will be used when linking the persons to the *Realencyclopädie*.

The BIO standard format will be used for Named Entity Tagging.

2.1.3 Will search keywords be provided in the metadata to optimize the possibility for discovery and then potential re-use?

- Yes: describe below

2.1.4 Will metadata be offered in such a way that it can be harvested and indexed?

- Yes: describe below

The dataset will be published in a data repository, that will generate discovery metadata that can be harvested and indexed.

2.2 FAIR data: Making data accessible

2.2.1 Will the data and other research outputs be deposited in a trusted repository?

- Yes: describe below

Zenodo repository will be used.

2.2.2 Have you explored appropriate arrangements with the identified repository where your data and other research outputs will be deposited?

- Yes

2.2.3 Does the repository ensure that the data and other research outputs are assigned an identifier? Will the repository resolve the identifier to a digital object?

Yes, Zenodo uses DOIs, which resolve to the digital Zenodo repository, containing the digital dataset.

2.2.4 Will all data and other research outputs be made openly available?

- Yes

2.2.5 Is an embargo applied to give time to publish or seek protection of the intellectual property (e.g. patents)?

- No

2.2.6 If an embargo is applied (see question 2.2.5), specify why and how long this will apply, bearing in mind that research data should be made available as soon as possible.

Not applicable

2.2.7 Will the data and other research outputs be accessible through a free and standardized access protocol?

- Yes: describe below

Yes, the HTTP protocol.

2.2.8 If there are restrictions on use, how will access be provided to the data, both during and after the end of the project?

Not applicable

2.2.9 How will the identity of the person accessing the data be ascertained?

Not applicable

2.2.10 Is there a need for a data access committee (e.g. to evaluate/approve access requests to personal/sensitive data)?

- No

2.2.11 Will metadata be made openly available and licenced under a public domain dedication CC0, as per the Grant Agreement? If not, please clarify why.

- Yes

2.2.12 Will metadata contain information to enable the user to access the data?

- Yes

The README file will contain all the necessary information.

2.2.13 How long will the data remain available and findable? Will metadata be guaranteed to remain available after data is no longer available?

The data will remain available and findable as long as Zenodo exist.

2.2.14 Will documentation or reference about any software needed to access or read the data be included? Will it be possible to include the relevant software (e.g. in open source code)?

The data will be accessible without the need of a special software.

2.3 FAIR data: Making data interoperable

2.3.1

What data and metadata vocabularies, standards, formats or methodologies will you follow to make your data interoperable to allow data exchange and re-use within and across disciplines? Will you follow community-endorsed interoperability best practices? Which ones?

The data files will be in published in standard and commonly used formats.

When linked to the LiLa Knowledge Base, standard vocabularies will be used (Ontolex-Lemon, dcterms...)

2.3.2 In case it is unavoidable that you use uncommon or generate project specific ontologies or vocabularies: Will you provide mappings to more commonly used ontologies? Will you openly publish the generated ontologies or vocabularies to allow reusing, refining or extending them?

If the modelling and linking to LiLa will require to generate project specific ontologies, they will be mapped to the most commonly used ones and they will be openly published.

2.3.3 Will your data and other research outputs include qualified references to other data (e.g. other data from your project, or datasets from previous research)?

- Yes

The annotated dataset will include a qualified reference to the original dataset.

2.4 FAIR data: Increase data re-use

2.4.1 How will you provide documentation needed to validate data analysis and facilitate data re-use?

The repository will contain a README file that will explain how the data was generated and any other relevant information about how to use it.

2.4.2

Will your data and other research outputs be made freely available in the public domain to permit the widest re-use possible? Will your data and other research outputs be licensed using standard reuse licenses, in line with the obligations set out in the Grant Agreement?

Yes.

2.4.3 Will the data and other research output produced in the project be useable by third parties, in particular after the end of the project?

- Yes

2.4.4 Will the provenance of the data and other research outputs be thoroughly documented using the appropriate standards?

- Yes

2.4.5 Describe all relevant data quality assurance processes.

Guidelines will be defined before starting the annotation process. Each step of the process will be documented.

3. Other research outputs

3.1 Do you have any additional information, that was not addressed in the previous sections, which you wish to provide regarding other research outputs that are generated or re-used throughout the project?

No

4. Allocation of resources

4.1 What will the costs be for making data and other research outputs FAIR in your project?

There might be some costs related to the publication in Open Access journals, only if it is unavoidable.

4.2 How will these be covered?

With the European Grant.

4.3 Who will be responsible for data management in your project?

Me, Valeria Boano, with the supervision of Margherita Fantoli and Monica Berti.

4.4 How will long term preservation be ensured?

As long as Zenodo exists.

5. Data security

5.1 What provisions are or will be in place for data security?

The data will be stored on Microsoft Onedrive, so that a backup is always available.

5.2 Will the data be safely stored in trusted repositories for long term preservation and curation?

- Yes

On Zenodo.

6. Ethics

6.1 Are there, or could there be, any ethics or legal issues that can have an impact on data sharing?

- No

6.2 Will informed consent for data sharing and long term preservation be included in questionnaires dealing with personal data?

- Not applicable

7. Other issues


7.1 Do you, or will you, make use of other national/funder/sectorial/departmental procedures for data management? If yes, which ones (please list and briefly describe them)?

- No

2. The philosophical canon and the art of (mis)quoting Plato and Aristotle in the CAG

DATA MANAGEMENT PLAN

(To be filled in and uploaded as deliverable in the Portal Grant Management System, at the due date foreseen in the system (and regularly updated).)

 *The template is recommended but not mandatory. If you do not use it, please make however sure that you comply with the research data management requirements under Article 17 of the Grant Agreement.)*

PROJECT	
Project number:	[101120349]
Project acronym:	[MECANO]
Project name:	[The Mechanics of Canon Formation and the Transmission of Knowledge from Graeco-Roman Antiquity]

DATA MANAGEMENT PLAN	
Date:	[06/02/2025]
Version:	[DMP version 1]

Data Summary

Will you re-use any existing data and what will you re-use it for? State the reasons if re-use of any existing data has been considered but discarded.

I will re-use some existing data. I will be re-using the some of the data of the Open Greek and Latin Project of the University of Leipzig.¹ This project has made OCR's of my corpus, the *Commentaria in Aristotelem Graeca* (CAG) and the texts of Plato and Aristotle in the XML format. I will firstly converting these files to plain TXT files, cleaning up as many computer mistakes as possible and analysing them for text re-use within the framework of my project.

Other data I will be re-using is the existing database of the University of Muenster.² This database has compiled quotes of Plato by other authors. I will check this database and compare it to my own findings. I will therefore incorporate the findings of the University of Muenster in my research.

I am also planning to re-use one or multiple text re-use detection (TRD) tools. The tools I have in mind are PASSIM and a tool in construction by the Belgian publisher Brepols. I will re-use these tools to find text re-use, so that I can

¹ <https://www.opengreekandlatin.org/>, <https://scaife.perseus.org/library/>.

² <https://www1.ivv1.uni-muenster.de/litw3/platon/indexP01.htm>

speak of the canonization of Aristotelian and Platonic philosophies on the basis of the text re-use.

Finally, I will be reading papers and books on Neoplatonism, the authors of the CAG, quotations and TRD. If helpful, I will re-use techniques and approaches while crediting the authors.

What types and formats of data will the project generate or re-use?

This project will generate and re-use a number of different types of data. The data that I re-use will be XML files, which will be transformed into TXT files. I will also re-use the data on the website of the University of Muenster regarding the quotes of Plato. I will also re-use some TRD tools, and henceforth the coding of these tools, in this project.

The data I will generate will also be of different kinds. I will generate a database of the quotations with MySQL in the program Filemaker. I am considering to also convert these databases into XML afterwards. I am furthermore considering to convert my database finding into HTML so that I can make the database more easily accessible. I will also be writing articles and a doctoral dissertation for this project, of which the type of data will be text.

What is the purpose of the data generation or re-use and its relation to the objectives of the project?

The purpose of the data generation or re-use is the finding of quotations by Late Antique and early Byzantine commentators of Aristotle and Plato. The objective of the project is to discuss the canonicity of Plato and Aristotle, and specifically which works of Plato and Aristotle were canonical, which is why the re-use of the XML files, earlier findings and suits like PASSIM are necessary to find the quotations.

What is the expected size of the data that you intend to generate or re-use?

The expected size of the data that I intend to generate is very much dependent on how many of my case studies I will be able to accomplish. I expect, if I were to be able to analyse every commentary of the corpus of the CAG, that I will end up with a database which would contain roughly 25.000 quotes. I base this upon the findings of Mirjam Kotwick (2016), who found 1000 quotes of Aristotle in one volume. As there are 23 volumes and also quotes of Plato, I think that the number of 25.000 is reasonable. If I just look at my first case study, this is roughly 4 volumes, of which one is in Latin and Hebrew, which will turn up roughly 5.000-6.000 quotes based upon earlier reasoning.

The expected size of the re-used data is much larger, as this contains every work of Plato, Aristotle and the CAG in XML format. I will need to reformat this to TXT and make it a good running text. Then there is also the suits of which I will re-use data and the data which was already found to be able to cross-reference with my own database. In terms of datasize such as how many kb/mb/gb, the size will not be extremely big as the XML files are not very heavy. It is a different case when speaking about characters. The medium sized files in the CAG contain roughly 1 million ancient Greek characters and a volume normally is made up from either one large commentary, two medium sized commentaries, or several smaller commentaries. This would mean that I will end up with roughly 50 million characters of ancient Greek if we were to only talk about the CAG and ignore the Aristotelian and Platonic corpora.

What is the origin/provenance of the data, either generated or re-used?

The origin of the data could all be placed with either Plato, Aristotle, the commentators and the editors of the original volumes. The only data that has another origin than this is the data of the suits meant to find the quotations. The origin of these suits is David Smith in the case of PASSIM or BREPOLS in the case of their text re-use detection suite, which is still in construction as of now.

To whom might your data be useful ('data utility'), outside your project?

My data will first and foremost be useful for other scholars outside my project. The first group that will be able to use my data will be other philosophers. They will be able to do their own research on the history of philosophy and the forming of a canon on the basis of the data that I use. Furthermore, linguists will be able to use the cleaned txt files to lemmatize the text and do research on words use and other phenomena. Finally, Digital humanists might find my data useful to replicate similar experiments with other data to see what kind of results they find and how they compare/differ.

FAIR data

Making data findable, including provisions for metadata

Will data be identified by a persistent identifier?

Data will be identified by PID for this project. I will assign a DOI to my database which will make it publicly accessible. Furthermore, I will upload the database to a data repository, although I am not sure which one as of this moment. I will also provide DOIs for my articles and Dissertation and make sure that they are open-access. For the XML and TXT data outputs, I will aim to have at least DOIs, but if KU Leuven supports them, I will see if PURLs or ARKs are possible and necessary. I will upload the source code that I generate on either Github or GitLab and the code that I re-use, I will reference rightly. For further re-use I will cite the datasets from the Open Greek and Latin Project and the University of Münster using their existing PIDs, such as DOIs, ARKs, or URNs, if they are provided. If they don't have PIDs, I will include stable URLs and versioning information in my references. If I indeed choose to use my MySQL database for a HTML valorisation project, I will also provide DOIs or similar PIDs.

Will rich metadata be provided to allow discovery? What metadata will be created? What disciplinary or general standards will be followed? In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how.

Rich metadata will be provided to ensure that all data generated and re-used in this project is discoverable and reusable. Metadata will describe the datasets, tools, and outputs of the project comprehensively, adhering to both general and disciplinary standards where available. I see seven possible categories of metadata that will most likely be created while working on this project: Descriptive Metadata, Administrative Metadata, Technical Metadata, Provenance Metadata, Structural Metadata, Disciplinary Metadata and Metadata for Publications.

1. **Descriptive Metadata**

- Dataset title, abstract/description, creator(s), keywords, publication date, and persistent identifiers (e.g., DOI).
- These elements will help users identify and understand the purpose of the data.

2. **Administrative Metadata**

- Information on access rights (e.g., open access or embargo), licensing terms (e.g., Creative Commons licenses), and contact details for the research team.

3. **Technical Metadata**

- File formats (e.g., TXT, XML, SQL, HTML), file sizes, software requirements (e.g., MySQL, FileMaker, TRD tools).
- This will ensure that users understand the technical requirements for accessing and using the data.

4. **Provenance Metadata**

- Documentation of the origins of the data (e.g., source databases like the Open Greek and Latin Project and the University of Münster's database).
- Details of processing steps, including cleaning, transforming, and analyzing data.
- Version history of datasets and tools.

5. **Structural Metadata**

- Organization of data, such as database schema and relationships between datasets (e.g., linking cleaned TXT files to database entries).

6. **Disciplinary Metadata**

- Use of controlled vocabularies, citation standards, and ontologies relevant to classical philosophy and text reuse.

7. **Metadata for Publications**

- Bibliographic information for articles and dissertations, including title, authors, journal name, and associated datasets.

I will aim to follow the general standards which are set out in the FAIR principles and the disciplinary standards for XML, the TEI.

In areas where metadata standards are lacking for classical philosophy and text reuse studies, custom metadata will be created with the following components: **Content Description** (Titles, authors, and abstracts for textual content), **Processing History** (Detailed steps of data transformation and analysis) and **Relationships** (Explicit links between datasets, tools, and research outputs). This approach will ensure that the data and outputs of this project are well-documented, discoverable, and reusable across disciplines.

Will search keywords be provided in the metadata to optimize the possibility for discovery and then potential re-use?

Yes, search keywords will be included in the metadata to optimize the discoverability and potential re-use of the project's data and outputs. These keywords will be carefully selected to reflect the core themes and methodologies of the research. They will include terms relevant to classical philosophy, such as "Plato," "Aristotle," "Neoplatonism," and "philosophical canon," as well as methodological keywords like "text reuse," "commentaries," "quotations," and "text reuse detection tools."

Additionally, keywords will be chosen to align with established vocabularies and ontologies in classical studies and digital humanities to enhance compatibility with existing discovery platforms. This approach will ensure that users from various academic and non-academic backgrounds can easily locate and engage with the project's outputs, thereby maximizing their impact and re-use potential.

Will metadata be offered in such a way that it can be harvested and indexed?

Yes, the metadata for this project will be structured and offered in a format that facilitates harvesting and indexing by discovery platforms and search engines. Metadata will comply with widely recognized standards such as FAIR for general metadata and TEI (Text Encoding Initiative) for textual data in XML format. This ensures compatibility with metadata aggregation services and academic repositories.

The metadata will include persistent identifiers (e.g., DOIs) and be hosted in open-access repositories or institutional platforms that support automated harvesting protocols like OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting). By adhering to these standards and practices, the project's metadata will be easily indexed by services like Zenodo or institutional discovery systems, significantly enhancing its accessibility and visibility to researchers and other users.

Making data accessible

Repository:

Will the data be deposited in a trusted repository?

Yes, the data will be deposited in a trusted repository. This will either be an institutional repository system or another trusted repository system such as Zenodo. This is not chosen as of now but will be more concrete in the coming period.

Have you explored appropriate arrangements with the identified repository where your data will be deposited?

At this stage, I have not finalized arrangements with a specific repository for data deposition. However, I am actively exploring suitable options to ensure the data is stored in an appropriate and accessible manner.

Does the repository ensure that the data is assigned an identifier? Will the repository resolve the identifier to a digital object?

At this stage, I have not finalized arrangements with a specific repository for data deposition. However, I am actively exploring suitable options to ensure the data is stored in an appropriate and accessible manner.

Data:

Will all data be made openly available? If certain datasets cannot be shared (or need to be shared under restricted access conditions), explain why, clearly separating legal and contractual reasons from intentional restrictions. Note that in multi-beneficiary projects it is also possible for specific beneficiaries to keep their data closed if opening their data goes against their legitimate interests or other constraints as per the Grant Agreement.

Yes, I aim to make all data generated through this project openly available to the fullest extent possible, adhering to the principles of transparency, reproducibility, and the FAIR data principles (Findable, Accessible, Interoperable, and Reusable). Open access to data is a cornerstone for fostering collaboration, innovation, and ensuring the broader societal impact of research outputs. There does not seem to be any problem with restricted

access at this moment.

If an embargo is applied to give time to publish or seek protection of the intellectual property (e.g. patents), specify why and how long this will apply, bearing in mind that research data should be made available as soon as possible.

There is nothing that I want to publish as of now which would have an embargo.

Will the data be accessible through a free and standardized access protocol?

Yes, all data generated through this project will be made accessible through a free and standardized access protocol. Ensuring accessibility is a key priority to maximize the usability and impact of the data.

If there are restrictions on use, how will access be provided to the data, both during and after the end of the project?

There are no restrictions on use

How will the identity of the person accessing the data be ascertained?

The identity of individuals accessing the data will depend on the access policies applied to different types of outputs. For data intended for open access, such as processed datasets, publications, and metadata, no authentication will be required, as these will be freely available to maximize visibility and reuse.

For any data or tools that might have restricted access due to licensing agreements, ethical considerations, or embargo periods, identity verification will be managed through the hosting platform or repository. Institutional repositories and platforms like Zenodo or institutional repositories typically require users to log in with institutional credentials or personal accounts to access restricted materials.

If necessary, additional measures such as access requests or the use of secure data-sharing systems could be implemented. These mechanisms ensure that sensitive or restricted data is accessible only to authorized individuals while maintaining the principles of accountability and secure data sharing.

Is there a need for a data access committee (e.g. to evaluate/approve access requests to personal/sensitive data)?

I do not see the necessity for a data access committee as of now as I do not believe that the data collected, used or re-used is personal or sensitive for my project. If this were to change, I would come back to this point.

Metadata:

Will metadata be made openly available and licenced under a public domain dedication CC0, as per the Grant Agreement? If not, please clarify why. Will metadata contain information to enable the user to access the data?

Yes, the metadata will be made openly available and licenced under a public domain dedication CC0, as per the Grant Agreement. The metadata will also contain information to enable the user to access the data.

How long will the data remain available and findable? Will metadata be guaranteed to remain available after data is no longer available?

The data in the repositories should be available for a long time as this is dependant on the repositories themselves. The HTML application I have in might have a shorter availability, as it is up to the hosting. The data behind the HTML and also the database itself will also be stored in the repository.

Will documentation or reference about any software be needed to access or read the data be included? Will it be possible to include the relevant software (e.g. in open source code)?

Documentation or reference to any software to access or read the data should not be necessary, but if people would like to replicate the research, reference to the software and their documentation should be made. It will not be able to include the relevant software, as it is owned by others, but it is open access, so reference to the relevant software should be sufficient.

Making data interoperable

What data and metadata vocabularies, standards, formats or methodologies will you follow to make your data

interoperable to allow data exchange and re-use within and across disciplines? Will you follow community-endorsed interoperability best practices? Which ones?

To ensure the interoperability of the project's data and metadata and facilitate data exchange and reuse within and across disciplines, recognized vocabularies, standards, formats, and methodologies will be employed throughout the research.

For metadata, the project will adhere to community-endorsed standards such as **Dublin Core** for general-purpose metadata, ensuring compatibility with a wide range of discovery and repository platforms. Textual data will follow the **TEI (Text Encoding Initiative)** guidelines, a widely recognized standard in the humanities for encoding textual resources in XML. These standards ensure that the metadata and data are structured in a way that supports seamless integration with other datasets and tools.

The data formats will include **TXT** for cleaned and processed text files, **XML** for encoded textual data, and **SQL** for databases. Where possible, these formats will be accompanied by conversion options (e.g., exporting the database into HTML) to enhance usability.

In terms of interoperability best practices, the project will align with the **FAIR Principles (Findable, Accessible, Interoperable, Reusable)**. Persistent identifiers (e.g., DOIs) will be assigned to datasets to ensure their stable and unique identification. Controlled vocabularies and ontologies relevant to classical studies and text reuse will be used to align with disciplinary norms. For example, terminologies related to Platonic and Aristotelian philosophy, as well as digital humanities methodologies like text reuse detection, will be incorporated.

The project will also adopt open and widely supported data-sharing protocols, where applicable, to ensure data and metadata can be easily harvested and integrated into other systems. By following these standards and best practices, the project aims to maximize the interoperability and reusability of its outputs across disciplines and platforms.

In case it is unavoidable that you use uncommon or generate project specific ontologies or vocabularies, will you provide mappings to more commonly used ontologies? Will you openly publish the generated ontologies or vocabularies to allow reusing, refining or extending them?

In the event that project-specific or uncommon ontologies or vocabularies are developed to support the research, the project will provide clear mappings to more widely recognized and commonly used ontologies wherever possible. This will ensure that the data remains interoperable with other datasets and systems in the broader research community. For instance, any custom vocabularies related to specific philosophical terms or text reuse patterns will be mapped to established ontologies used in classical studies, digital humanities, or the broader field of philosophy.

Moreover, the project will openly publish any generated ontologies or vocabularies to enable others to reuse, refine, and extend them. These will be made available through open-access repositories such as Zenodo or GitHub, and will include clear documentation to facilitate their adoption and further development by other scholars. By doing so, we aim to contribute to the broader academic ecosystem and foster collaboration and innovation in the field. This approach aligns with best practices for open science and ensures that the resources developed during the project can be used to benefit future research initiatives.

Will your data include qualified references³ to other data (e.g. other data from your project, or datasets from previous research)?

Yes, the data generated in this project will include qualified references to other data, both from within the project itself and from external datasets. This will help establish connections between different components of the research, as well as to relevant resources from prior research. For instance, the project will include references to previous works and datasets, such as the Open Greek and Latin Project and the University of Münster's database of Plato quotations, which will be incorporated into the analysis. These references will be made explicit in the metadata and data files, ensuring that relationships between different data sources are clearly documented.

Additionally, within the project itself, various datasets (e.g., the database of quotations, the processed text files, and any results from text reuse detection) will be interlinked, with references to one another clearly outlined.

³ A qualified reference is a cross-reference that explains its intent. For example, *X is regulator of Y* is a much more qualified reference than *X is associated with Y*, or *X see also Y*. The goal therefore is to create as many meaningful links as possible between (meta)data resources to enrich the contextual knowledge about the data. (Source: <https://www.go-fair.org/fair-principles/i3-metadata-include-qualified-references-metadata/>)

This will allow users to trace how different data points relate to each other and how they contribute to the research. These references will be made using standard citation practices and linked through persistent identifiers (such as DOIs or URNs) to facilitate tracking and ensure the reproducibility and reusability of the research.

By including these qualified references, the project aims to provide a comprehensive, transparent, and interconnected dataset that enables future researchers to explore related datasets and build upon the work conducted.

Increase data re-use

How will you provide documentation needed to validate data analysis and facilitate data re-use (e.g. readme files with information on methodology, codebooks, data cleaning, analyses, variable definitions, units of measurement, etc.)?

To validate the data analysis and facilitate data reuse, comprehensive documentation will be provided for all datasets and methodologies used in the project. This will include clear and detailed **readme files** that describe the entire data processing pipeline, from the raw data sources through to the final outputs. The readme files will include information on the **methodology** employed, such as how the text reuse detection tools (TRD) were applied, how data was cleaned, and the analytical frameworks used for interpreting quotations and citations of Plato and Aristotle in the commentaries.

Additionally, the project will provide **codebooks** that outline the structure of the datasets, including definitions of variables, the criteria for text reuse detection, and any custom ontologies or vocabularies created during the research. This will ensure that users understand how the data was classified and processed. Where applicable, the codebooks will include explanations of **units of measurement** and **categorizations** used to analyze quotations, sources, and other key data points.

The documentation will also cover the **data cleaning** process, detailing how OCR errors were corrected, how the raw data was transformed from XML into TXT files, and any additional steps taken to ensure the accuracy and reliability of the final datasets.

By making these documents publicly available alongside the datasets and ensuring they are clearly structured and easy to follow, the project will support both the validation of the research process and the seamless reuse of the data by other scholars. This approach aligns with best practices for transparency in data science and ensures that others can reproduce, refine, and build upon the work conducted.

Will your data be made freely available in the public domain to permit the widest re-use possible? Will your data be licensed using standard reuse licenses, in line with the obligations set out in the Grant Agreement?

Yes, the data will be made freely available in the public domain to permit the widest re-use possible. My data will also be licensed using standard reuse licenses, in line with the obligations set out in the Grant Agreement.

Will the data produced in the project be useable by third parties, in particular after the end of the project?

Yes, the data produced in the project will be useable by third parties, also after the end of the project.

Will the provenance of the data be thoroughly documented using the appropriate standards?

Yes, the provenance of the data will be thoroughly documented using the appropriate standards

Describe all relevant data quality assurance processes.

To ensure the highest standards of data quality throughout the project, a robust set of quality assurance processes will be implemented at each stage of the data lifecycle. These processes will address data accuracy, consistency, completeness, and reliability.

Data Collection and Preprocessing

The project will begin by re-using existing data from trusted sources, such as the Open Greek and Latin Project and the University of Münster's Plato quotations database. These sources have already undergone rigorous quality checks, and their reliability will be further verified during the preprocessing phase. During this phase, data will be converted from XML to TXT format, and any OCR errors or inconsistencies will be identified and corrected. This process will be carried out manually and programmatically, with multiple rounds of validation to ensure

accuracy.

Text Reuse Detection (TRD)

For the text reuse detection process, well-established tools like PASSIM and Brepols' in-development tool will be used. These tools will undergo thorough testing and calibration to ensure that they produce reliable results when applied to the project's datasets. We will also perform cross-validation by comparing the results from the TRD tools with manually identified quotations, ensuring that the automated detection aligns with human judgment. Any discrepancies will be carefully examined and addressed to refine the accuracy of the text reuse analysis.

Data Analysis and Documentation

During the analysis phase, careful attention will be paid to defining variables, categorizing quotations, and analyzing relationships between texts. A detailed **codebook** will be created to outline the methodology, definitions, and categories used, ensuring that all analyses are transparent and reproducible. Additionally, data cleaning processes will be documented to track any modifications made to the raw data.

Quality Checks for Outputs

The final datasets, such as the MySQL database of quotations and any transformed outputs (e.g., HTML or XML exports), will undergo a final round of validation to ensure that all data points are consistent and accurate. This includes verifying that the correct texts and quotations are linked, the relationships between them are properly documented, and all relevant metadata (e.g., persistent identifiers, keywords) is included.

Continuous Documentation and Feedback

Throughout the project, detailed **readme files** and **methodology documentation** will be maintained, providing an ongoing record of data processing steps, quality assurance measures, and any issues encountered. These documents will be made publicly available alongside the datasets, ensuring transparency and allowing others to assess and replicate the quality assurance process.

By employing these rigorous quality assurance processes at every stage of the project, we aim to ensure that the data produced is accurate, reliable, and of high scholarly value, contributing to the integrity of the research and its potential for reuse by other scholars.

Further to the FAIR principles, DMPs should also address research outputs other than data, and should carefully consider aspects related to the allocation of resources, data security and ethical aspects.

Other research outputs

In addition to the management of data, beneficiaries should also consider and plan for the management of other research outputs that may be generated or re-used throughout their projects. Such outputs can be either digital (e.g. software, workflows, protocols, models, etc.) or physical (e.g. new materials, antibodies, reagents, samples, etc.).

Beneficiaries should consider which of the questions pertaining to FAIR data above, can apply to the management of other research outputs, and should strive to provide sufficient detail on how their research outputs will be managed and shared, or made available for re-use, in line with the FAIR principles.

Allocation of resources

What will the costs be for making data or other research outputs FAIR in your project (e.g. direct and indirect costs related to storage, archiving, re-use, security, etc.) ?

There will be some costs for the hosting of the database, which will be hosted by ARTES, a digital humanities institution of KU Leuven. Artes is willing to host, provide access to filemaker and to give access to work and change the database to all people involved in the project and will host the database. They ask for roughly 150 euros per year.

How will these be covered? Note that costs related to research data/output management are eligible as part of the Horizon Europe grant (if compliant with the Grant Agreement conditions)

The costs will most likely be covered by the bench fee which in turn is provided by the grant agreement

Who will be responsible for data management in your project?

I will be responsible first and foremost for the data management and ARTES and KU Leuven will be responsible for the hosting of the database.

How will long term preservation be ensured? Discuss the necessary resources to accomplish this (costs and potential value, who decides and how, what data will be kept and for how long)?

Long term preservation will be ensured by the uploading of the data in trusted repositories while they have all the necessary PIDs. Only the database will not be able to be hosted for a long time after the end of the project, but I will also upload the database in the repository, either in XML format or in MySQL format.

Data security

What provisions are or will be in place for data security (including data recovery as well as secure storage/archiving and transfer of sensitive data)?

Ensuring the security of the data throughout the project's lifecycle is a priority. Provisions will be made for secure storage, transfer, and recovery of all data, in accordance with best practices and relevant institutional policies. All data, including both raw and processed materials, will be stored securely using reliable platforms that comply with relevant security standards. For sensitive data, such as unpublished research results or data with restrictions on access, secure storage solutions with encryption and access control will be utilized. These may include institutional repositories or trusted third-party platforms like Zenodo or GitHub for software, ensuring proper user authentication. For long-term storage and archiving, we will use open-access, trusted archival services that comply with relevant standards (e.g., LOCKSS, CLOCKSS), guaranteeing both the security and future accessibility of the data. When transferring sensitive data, encryption protocols (e.g., HTTPS, SSH) will be used to ensure that data remains secure during transit. Access control measures, such as user authentication, role-based access permissions, and logging, will be put in place to limit access to sensitive or restricted data to authorized users only. For data sharing with collaborators or other stakeholders, secure file-sharing platforms will be employed to ensure the protection of sensitive information.

To mitigate any risk of data loss, a comprehensive data recovery plan will be implemented. This plan will include regular backups of all project data, stored in geographically distributed locations to prevent loss due to hardware failures or other emergencies. Backup systems will be regularly tested to ensure that data can be recovered quickly and accurately in the event of an incident. In the unlikely event that sensitive personal data, proprietary content, or confidential information is involved in the project, additional provisions will be made to ensure compliance with relevant data protection regulations (e.g., GDPR). Sensitive data will be anonymized or pseudonymized wherever possible, and access to such data will be strictly limited to authorized personnel. Detailed data management and security protocols will be developed to ensure that all sensitive data is handled with the utmost care and in accordance with legal and ethical guidelines.

By implementing these provisions, the project will ensure the integrity, security, and availability of its data, supporting both its short-term research goals and long-term sustainability, while ensuring compliance with relevant data protection and privacy standards.

Will the data be safely stored in trusted repositories for long term preservation and curation?

Yes, the data will be safely stored in trusted repositories for long term preservation and curation.

Ethics

Are there, or could there be, any ethics or legal issues that can have an impact on data sharing? These can also be discussed in the context of the ethics review. If relevant, include references to ethics deliverables and ethics chapter in the Description of the Action (DoA).

There should not be any ethical or legal issues which have an impact on data sharing.

Will informed consent for data sharing and long term preservation be included in questionnaires dealing with personal data?

There will be no usage of personal data of others in this research project.

Other issues

Do you, or will you, make use of other national/funder/sectorial/departmental procedures for data management? If yes, which ones (please list and briefly describe them)?

HISTORY OF CHANGES		
VERSIO N	PUBLICATION DATE	CHANGE
1.0	06-02-2025	Initial version (new MFF).

3. The presence of classics in early modern book history - Data management plan (after 6 months)

Plan Overview

A Data Management Plan created using DMPTuuli

Title: The presence of classics in early modern book history

Creator: Jonas Paul Fischer

Principal Investigator: Jonas Paul Fischer

Affiliation: University of Helsinki

Funder: European Commission

Template: Horizon Europe - Data Management Plan

ORCID iD: 0009-0008-6283-3637

Project abstract:

The aim of this PhD project is to study the printing of Graeco-Roman texts and their reuse in Early Modern Britain. The Computational History Group Helsinki (COMHIS) has gathered a vast dataset about textual overlaps within the Eighteenth Century Collection Online (ECCO) and has compiled and enriched and cleaned version of the English Short Title Catalog (ESTC). The present PhD project builds upon this prior work and uses the ESTC metadata to analyze how the print and translation industry engaged with the Graeco-Roman classics in 16th to 18th century Britain. The overlap data for ECCO will be used to study quotes from Graeco-Roman texts in 18th century books in order to shed some light on quotation practices, language use and the reception of the classics.

ID: 25895

Start date: 01-09-2024

End date: 31-08-2028

Last modified: 04-02-2025

Grant number / URL: 101120349

1. Data Summary

Date of the DMP

04.02.2025

DMP version

1.0

1. Data Summary

I will use existing data provided by the English Short Title Catalog (ESTC) and Eighteenth Century Collection Online (ECCO) databases as well as data about these databases provided to me by the Computational History Group Helsinki (COMHIS). I will reuse and enrich this data for my analysis and to answer new research questions. ESTC metadata will be used to study the printing and translation industry that worked with the Graeco-Roman Classics in Early-Modern Britain. ECCO data will be used to systematically find overlap between the Classics and other printed material in 18th century Britain in order to analyze citation practices.

Data used in this project is in CSV and PDF formats. Data generated in this project will be in CSV format. Since CSV files are rather economic in their need for storage space, I expect the size of the data generated to be negligible.

Data generated in this project might be useful anyone doing research on the reception of the Graeco-Roman Classics in Early-Modern Britain.

2. FAIR data

2.1. Making data findable, including provisions for metadata

Data will receive a persistent identifier, permanent URL and be connected to my ORCID ID.

Metadata and search keywords will be provided.

2.2. Making data accessible

Data will be deposited in trusted repository. Arrangements have not been made yet, though Fairdata generally provides storage space to all Finnish research institutions, making a separate arrangement unnecessary. Fairdata will assign an identifier.

The data I create during the project and that is based on the ESTC and ECCO data will be openly shared. The underlying data from ESTC and ECCO can not be openly shared though, since these databases are operated by commercial companies that do not allow the open sharing of their data. ESTC and ECCO are available through their respective owners though.

Since this project does not use or produce any sensitive data an access committee or other methods or identifying the person accessing the data are not necessary.

Data will be shared in CSV format, making any special software unnecessary.

2.3. Making data interoperable

Data will be shared in CSV format, allowing for easy interoperability and reuse. No special vocabulary, standards, etc. will be created.

Qualified references to other data will be provided.

2.4. Increase data re-use

Readme files with descriptions of the data creation process (including methodology, data cleaning, data provenance, etc.) will be provided.

Data will be made freely available and useable for third parties.

3. Other research outputs

3. Other research outputs

Digital workflows will be saved in my Github and be open to the public.

4. Allocation of resources

4. Allocation of resources

There are no costs for making data or research outputs for my project FAIR. By storing any data I can legally share in the Fairdata digital repository, everything will be open for anyone to access. Since this repository provides within certain limits free storage space to Finnish educational institutions, there are no direct costs for this project and the operator will ensure long term preservation, accessibility and security.

I will be responsible for data management myself.

Long term data preservation will be ensured by storing it in a trusted repository like the Fairdata digital repository.

5. Data security

5. Data security

Since there is no especially sensitive data in this project no special measures have to be taken to ensure access security. In the short-term data storage and recovery will be ensured by saving all materials in trusted repositories and a local device at the same time. In the long term data will be stored in a trusted repository like the Fairdata digital repository for long term preservation.

6. Ethics

6. Ethics

There are no ethics issues to be considered here. Since the project will be dealing with books from early modern times, the safekeeping of personal information does not need to be considered.

Since the underlying data for my research is partly provided by third parties, they can not be directly shared. Instead their databases will be referenced.

7. Other issues


7. Other issues

I will make use of the University of Helsinki Computational History Group's internal data management procedures. This includes the storage of any newly created materials in the group's Github and Google Drive repositories and the documentation of processes.

4. Pulse and Physiology in Hellenistic Science

DATA MANAGEMENT PLAN

(To be filled in and uploaded as deliverable in the Portal Grant Management System, at the due date foreseen in the system (and regularly updated).)

 *The template is recommended but not mandatory. If you do not use it, please make however sure that you comply with the research data management requirements under Article 17 of the Grant Agreement.)*

PROJECT	
Project number:	101120349
Project acronym:	MECANO
Project name:	MECANO: The Mechanics of Canon Formation and the Transmission of Knowledge from Greco-Roman Antiquity

DATA MANAGEMENT PLAN	
Date:	16/2/2025
Version:	1.0

Data Summary

Will you re-use any existing data and what will you re-use it for? State the reasons if re-use of any existing data has been considered but discarded.

- Yes, for my research, I plan to reuse existing data from critical editions of ancient texts.

What types and formats of data will the project generate or re-use?

- The project will primarily generate and reuse textual data, bibliographic databases, and metadata related to sources relevant to my research.

What is the purpose of the data generation or re-use and its relation to the objectives of the project?

- The primary goal of this research is to map Hellenistic theories on the pulse, which have survived exclusively through indirect tradition – either as testimonies or fragments – due to their reception by later authors, particularly Galen. Galen devotes a significant portion of his extensive corpus to the study of the pulse, making his works essential for reconstructing earlier theories. The aim is to connect the textual analysis of the generated or reused data to the formation of the Hellenistic medical canon and its subsequent reception by Galen and other key figures, such as Rufus of Ephesus. The research will also involve two digital components:
 1. It will include the development of an online interface for visualizing and studying ancient sphygmology from both a conceptual and terminological perspective, likely through lexical entries.

2. It will involve the creation of a dataset of texts that will be automatically annotated using predefined labels. These annotations will ultimately require manual review and correction where necessary. The annotation process will be carried out using [INCEpTION](#), and the texts will likely be exported to [Coda](#) for further organization and structuring.

What is the expected size of the data that you intend to generate or re-use?

- The estimated size of the data can be determined using the TLG statistics tool, which provides word counts for each treatise. Since the study is based on a corpus rather than isolated fragments, the relevant dataset consists of Galen's works on the pulse, amounting to approximately 168,610 words. This includes *On the Function of the Pulse*, *On the Pulse of Beginners*, *Differences of Pulses*, *Diagnosis by Pulses*, *Causes of Pulses*, *Prognosis by Pulses*, and *Synopsis on Pulses*. The analysis will rely on this corpus for textual comparison and interpretation.

What is the origin/provenance of the data, either generated or re-used?

- The data will be sourced primarily from critical editions of ancient texts, collections of fragments and testimonia, and relevant secondary sources. Whenever available, texts will be retrieved from established online databases. For works not accessible in digital format, OCR will be performed on printed editions to produce machine-readable versions of the texts. In the case of Galen's works, editions such as Kühn's, which are no longer under copyright, will be used for digitization.

To whom might your data be useful ('data utility'), outside your project?

- The data may be useful to classicists, philosophers, and historians with an interest in Greek medical and philosophical texts.

FAIR data

Making data findable, including provisions for metadata

Will data be identified by a persistent identifier?

- A portion of the data, including tagged datasets – such as fragments annotated with keywords – will be made accessible on [Zenodo](#), which provides a DOI to ensure long-term stability and accessibility. Additional data may also be assigned persistent identifiers as needed.

Will rich metadata be provided to allow discovery? What metadata will be created? What disciplinary or general standards will be followed? In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how.

- It has not yet been determined whether rich metadata will be provided. The metadata that may be created, specifically at the time of the dataset's publication, will primarily relate to the content and source of the data itself, enabling searches based on keywords. The applicable standard could follow citation practices in Classical Studies, which require references to authors and works using the standard critical edition when citing an already published text.

Will search keywords be provided in the metadata to optimize the possibility for discovery and then potential re-use?

- Search keywords may be included in the metadata to optimize content discovery and potential re-use.

Will metadata be offered in such a way that it can be harvested and indexed?

- Metadata may be offered in a way that allows for harvesting and indexing, but the exact implementation has not yet been fully defined. Some metadata may be made available through the URLs of webpages in the planned interface, where elements such as lemmas or source authors will be part of the URL structure.

Making data accessible

Repository:

Will the data be deposited in a trusted repository?

- It is being considered to deposit the data in Zenodo. The platform would ensure long-term accessibility and proper archiving of the data.

Have you explored appropriate arrangements with the identified repository where your data will be deposited?

- Preliminary considerations have been made regarding the use of Zenodo for data deposition. Further arrangements will be explored to ensure that the repository meets the project's specific needs.

Does the repository ensure that the data is assigned an identifier? Will the repository resolve the identifier to a digital object?

- Since Zenodo assigns a DOI (Digital Object Identifier) to deposited materials, this would provide a persistent and citable reference for the data. The DOI would also be resolvable to a digital object, ensuring stable access over time.

Data:

Will all data be made openly available? If certain datasets cannot be shared (or need to be shared under restricted access conditions), explain why, clearly separating legal and contractual reasons from intentional restrictions. Note that in multi-beneficiary projects it is also possible for specific beneficiaries to keep their data closed if opening their data goes against their legitimate interests or other constraints as per the Grant Agreement.

- At least the raw data will be made openly accessible. Since there are no copyrights on the texts (as Kühn's editions of Galen are no longer under copyright), I do not foresee any datasets being restricted from sharing.

If an embargo is applied to give time to publish or seek protection of the intellectual property (e.g. patents), specify why and how long this will apply, bearing in mind that research data should be made available as soon as possible.

- Currently, there is no need for an embargo on the data, as the texts are in the public domain.

Will the data be accessible through a free and standardized access protocol?

- Yes, the data will be made accessible through a free and standardized access protocol, such as Zenodo, ensuring that it is open and easily accessible.

If there are restrictions on use, how will access be provided to the data, both during and after the end of the project?

- Since the data is not subject to copyright restrictions, there are no expected barriers to access. However, if any usage restrictions were to be imposed, they would be clearly outlined and access would be controlled via a request process or restricted access settings.

How will the identity of the person accessing the data be ascertained?

- The identity of those accessing the data will be verified through the registration system of the data repository, such as Zenodo, where users may need to sign in before accessing the data.

Is there a need for a data access committee (e.g. to evaluate/approve access requests to personal/sensitive data)?

- There is no personal or sensitive data in the scope of this project, so a data access committee is not necessary.

Metadata:

Will metadata be made openly available and licenced under a public domain dedication CC0, as per the Grant Agreement? If not, please clarify why. Will metadata contain information to enable the user to access the data?

- Metadata will be openly available and licensed under a public domain dedication (CC0) as per the Grant Agreement. They will contain all necessary information for users to access and navigate the data.

How long will the data remain available and findable? Will metadata be guaranteed to remain available after data

is no longer available?

- The data will remain available and findable as long as it is accessible on Zenodo. Metadata will also remain available for as long as possible through the platform.

Will documentation or reference about any software be needed to access or read the data be included? Will it be possible to include the relevant software (e.g. in open source code)?

- No specific software is required to access or read the data.

Making data interoperable

What data and metadata vocabularies, standards, formats or methodologies will you follow to make your data interoperable to allow data exchange and re-use within and across disciplines? Will you follow community-endorsed interoperability best practices? Which ones?

- For the project, I plan to follow standard practices from the field of Classics, particularly in the use of abbreviations for authors and works, which are widely recognized in scholarly circles. This will ensure interoperability within the discipline.

In case it is unavoidable that you use uncommon or generate project specific ontologies or vocabularies, will you provide mappings to more commonly used ontologies? Will you openly publish the generated ontologies or vocabularies to allow reusing, refining or extending them?

- I am not planning to create new ontologies or vocabularies for the project. However, if there are any project-specific terms or internal standards developed, I will ensure that they are well-documented and provide clear justification for their use.

Will your data include qualified references to other data (e.g. other data from your project, or datasets from previous research)?

- At this stage, no specific plans have been made to include references to other datasets beyond the project's own data. However, if the need arises, I will incorporate relevant references to ensure that the data can be contextualized and related to other work within the field.

Increase data re-use

How will you provide documentation needed to validate data analysis and facilitate data re-use (e.g. readme files with information on methodology, codebooks, data cleaning, analyses, variable definitions, units of measurement, etc.)?

- The documentation needed to validate data analysis and facilitate data re-use will be provided through standardized lists of abbreviations and conventions for citing authors and works. The abbreviations for Galen's works will follow those established in *The Oxford Handbook of Galen* (2024) while other abbreviations will align with those commonly recognized in Classical Studies, such as those in *The Oxford Classical Dictionary*, which covers both Greek and Latin authors and texts. This information may be provided to users through a README file.

Will your data be made freely available in the public domain to permit the widest re-use possible? Will your data be licensed using standard reuse licenses, in line with the obligations set out in the Grant Agreement?

- The raw data will be made freely available in the public domain to allow the widest possible re-use. As for more elaborated datasets, their availability is yet to be determined. If applicable, standard reuse licenses will be considered in accordance with the obligations outlined in the Grant Agreement.

Will the data produced in the project be useable by third parties, in particular after the end of the project?

- Yes, third parties will be able to use the data produced in the project, especially after its completion. The raw data will be openly accessible, allowing for broad reuse in future research. For more elaborated datasets, the aim is to structure and format them in a way that ensures their long-term reusability.

Will the provenance of the data be thoroughly documented using the appropriate standards?

- Yes, the provenance of the data will be thoroughly documented following appropriate standards. References to primary sources will adhere to established citation practices in Classical Studies, including author and work abbreviations as found in *The Oxford Handbook of Galen* (2024) for Galenic texts and in widely recognized resources such as the *Oxford Classical Dictionary* for other ancient authors. Metadata would eventually indicate the source of each dataset, specifying whether it derives from critical editions or secondary literature.

Describe all relevant data quality assurance processes.

- If the Greek text is obtained through OCR, it will be systematically compared with the printed editions to ensure accurate text recognition. This verification process will help maintain the integrity of the textual data and minimize errors in the digital transcription. Additionally, the automatic annotation of texts performed using INCEpTION will be reviewed to ensure internal consistency within the dataset in the analysis of the texts.

Further to the FAIR principles, DMPs should also address research outputs other than data, and should carefully consider aspects related to the allocation of resources, data security and ethical aspects.

Other research outputs

In addition to the management of data, beneficiaries should also consider and plan for the management of other research outputs that may be generated or re-used throughout their projects. Such outputs can be either digital (e.g. software, workflows, protocols, models, etc.) or physical (e.g. new materials, antibodies, reagents, samples, etc.).

- The data will be accessible through a dedicated platform (an interface) focused on the texts on the pulse, which are the subject of the research. This platform will be developed in accordance with FAIR principles and sustainability considerations and is expected to be supported by the Digital Humanities Centre of the Hebrew University of Jerusalem.

Beneficiaries should consider which of the questions pertaining to FAIR data above, can apply to the management of other research outputs, and should strive to provide sufficient detail on how their research outputs will be managed and shared, or made available for re-use, in line with the FAIR principles.

Allocation of resources

What will the costs be for making data or other research outputs FAIR in your project (e.g. direct and indirect costs related to storage, archiving, re-use, security, etc.)?

- The costs for making the data FAIR will primarily involve the development and maintenance of the interactive digital interface, which is a key aspect of the project. While data storage on Zenodo is free, ensuring that the interface remains accessible and functional in the long term will incur some costs. Additional costs may include the Open Access publication of articles derived from the data studied in the project.

How will these be covered? Note that costs related to research data/output management are eligible as part of the Horizon Europe grant (if compliant with the Grant Agreement conditions)

- The necessary funds for covering these costs will be included in the Horizon Europe grant.

Who will be responsible for data management in your project?

- I will be responsible for data management in the project, alongside my supervisor, Dr. Orly Lewis (HUJI), and my co-supervisor, Prof. Jan Opsomer (KU Leuven).

How will long term preservation be ensured? Discuss the necessary resources to accomplish this (costs and potential value, who decides and how, what data will be kept and for how long)?

- The data published via Zenodo will remain accessible as long as the platform supports it and ensures long-term preservation. Regarding the digital interface, long-term preservation plans are still under consideration, as its sustainability will depend on the platform and resources available at the time.

Data security

What provisions are or will be in place for data security (including data recovery as well as secure storage/archiving and transfer of sensitive data)?

- Regular backups of data will be performed to safeguard against potential loss. The data will be stored in cloud drives to ensure long-term accessibility and further protection.

Will the data be safely stored in trusted repositories for long term preservation and curation?

- As mentioned earlier, the data will be stored in Zenodo.

Ethics

Are there, or could there be, any ethics or legal issues that can have an impact on data sharing? These can also be discussed in the context of the ethics review. If relevant, include references to ethics deliverables and ethics chapter in the Description of the Action (DoA).

- As the research primarily involves editing fragments of texts from Galen and other ancient medical sources, there are no expected ethical or legal issues that would impact data sharing. The texts are in the public domain, and the project does not involve personal data or any sensitive information. There is no need for informed consent or data privacy considerations related to the data being used in this project.

Will informed consent for data sharing and long term preservation be included in questionnaires dealing with personal data?

- Informed consent for data sharing and long-term preservation will not be required, as the project does not involve personal data or sensitive information.

Other issues

Do you, or will you, make use of other national/funder/sectorial/departmental procedures for data management? If yes, which ones (please list and briefly describe them)?

- No, I will not be making use of other national, funder, sectorial, or departmental procedures for data management. All data management will follow the guidelines and procedures outlined in this plan.

HISTORY OF CHANGES		
VERSION	PUBLICATION DATE	CHANGE
1.0	05.05.2021	Initial version (new MFF).
1.1	01.04.2022	Reformatted to align with other deliverables templates.

5. Detecting and Retrieving Lost Historians

Leonardo D'Addario, Leipzig University

DATA MANAGEMENT PLAN

PROJECT	
Project number:	[131101134]
Project acronym:	[MECANO PhD 5]
Project name:	[Detecting and Retrieving Lost Historians]

DATA MANAGEMENT PLAN	
Date:	[20/02/2025]
Version:	[1.0]

Data Summary

Will you re-use any existing data and what will you re-use it for?

I will re-use the XML text of Polybius' *The Histories*, available from the Perseus Digital Library, and the data of the Jacoby Online, the digital edition of *Die Fragmente der Griechischen Historiker*. These XML text of *The Histories* will be used to digitally extract relevant linguistic elements that Polybius employs when quoting other historiographers (e.g., names of the authors, titles of their works, and verbs introducing paraphrases or verbatim quotations). This extraction will be used to detect Polybius' quoting practice. The data of the Jacoby Online will be used to analyse how historiographers cited by Polybius are quoted by other sources in order to establish a comparison between different quoting practices.

What types and formats of data will the project generate or re-use?

This project will re-use:

- Structured texts in XML (Polybius' *The Histories* from Perseus Digital Library)
- Structured texts in Markdown (MD) formats (from Jacoby Online)

This project will generate:

- CSV, HTML, JSON files of Polybius' *The Histories*
- UIMA CAS JSON files of *The Histories* (for working in the web-based platform INCEption)
- Extracted datasets of quotations' linguistic elements in tabular (TSV) formats.

What is the purpose of the data generation or re-use and its relation to the objectives of the project?

The XML files of *The Histories* will be converted into CSV files. These will be parsed with a trained spaCy NER model for Ancient Greek, in order to automatically annotate Named Entities. The output will generate files containing annotated NEs and will be formatted in UIMA CAS JSON. The UIMA CAS JSON files will be uploaded in the web-based platform for semantic annotation INCEpTION to correct the automatic annotation if necessary, and to manually complete the annotation of other relevant linguistic elements (e.g., verbs introducing paraphrases or quotation of historiographers). Then, the extracted datasets of linguistic elements will be used to analyse Polybius' quoting practice, which will contribute to the study of ancient Greek historiography canonization and transmission.

What is the expected size of the data that you intend to generate or re-use?

The project will begin with approximately 2 GB of re-used data and generate at least 1 GB, possibly more.

What is the origin/provenance of the data, either generated or re-used?

XML text of *The Histories* has been downloaded from the Perseus Digital Library, which is Open Access (<https://github.com/PerseusDL/canonical-greekLit>). The access to the data of the Jacoby Online (<https://gitlab.com/brillpublishers/data/jo>) has been provided by Brill Publisher via a license agreement. New data will be generated through digital text processing.

To whom might your data be useful ('data utility'), outside your project?

- Scholars in Classical Studies, particularly those studying fragmentary historiography
- Researchers in Digital Humanities and Computational Philology
- Projects working on digital editions and linguistic annotation of ancient texts

FAIR data

Making data findable, including provisions for metadata

Will data be identified by a persistent identifier?

During the manual annotation in INCEpTION, CTS URNs and Wikidata IDs will be assigned to the ancient authors and works quoted in Polybius' text. This will enable the unique identification of these authors and works, ensuring persistent and standardized referencing. The final dataset will be assigned a DOI with Zenodo (<https://zenodo.org/>).

Will rich metadata be provided to allow discovery? What metadata will be created? What disciplinary or general standards will be followed? In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how.

Named Entity labels will be used to classify personal names and place names, therefore generating metadata for onomastics and historical geography. Metadata about authors and their works quoted in Polybius' text will be created using TLG, the Perseus Catalog, Wikidata, and the Linked Ancient Greek and Latin (LAGL) project. They will also include links to external databases such as Wikidata and the Perseus Catalog (<https://catalog.perseus.org/>). If metadata about authors and works quoted in Polybius' text do not exist, they will be created according to the principles of TLG, the Perseus Catalog, and Wikidata.

Will search keywords be provided in the metadata to optimize the possibility for discovery and then potential re-use?

Yes.

Will metadata be offered in such a way that it can be harvested and indexed?

Metadata will be published in formats compatible with open metadata aggregators and linked data repositories.

Making data accessible

Repository:

Will the data be deposited in a trusted repository?

Data will be deposited in a GitHub repository with a Zenodo DOI.

Have you explored appropriate arrangements with the identified repository where your data will be deposited?

Does the repository ensure that the data is assigned an identifier? Will the repository resolve the identifier to a digital object?

Yes.

Data:

Will all data be made openly available? If certain datasets cannot be shared (or need to be shared under restricted access conditions), explain why, clearly separating legal and contractual reasons from intentional restrictions. Note that in multi-beneficiary projects it is also possible for specific beneficiaries to keep their data closed if opening their data goes against their legitimate interests or other constraints as per the Grant Agreement.

Data generated from the XML text of Polybius, downloaded from the Perseus Digital Library, will be openly available. However, the availability of data potentially generated from Jacoby Online must first be discussed with Brill Publishers. Access to Jacoby Online is not open, and I obtained permission to use this data under a License Agreement between the editor and me.

If an embargo is applied to give time to publish or seek protection of the intellectual property (e.g. patents), specify why and how long this will apply, bearing in mind that research data should be made available as soon as possible.

Will the data be accessible through a free and standardized access protocol?

Yes.

If there are restrictions on use, how will access be provided to the data, both during and after the end of the project?

Access to the XML text of Polybius on the Perseus Digital Library is open. During the semantic annotation process, access to INCEpTION is restricted to me and my supervisor, who can log in through an institutional account. At the end of the project, the data generated from the annotation will be openly available. The data from Jacoby Online are not openly accessible; access has been granted exclusively to me under a License Agreement with the editor. Therefore, the availability of data potentially generated from Jacoby Online must first be discussed with Brill Publishers.

How will the identity of the person accessing the data be ascertained?

The identity of the person accessing the data will be ascertained through an institutional login.

Is there a need for a data access committee (e.g. to evaluate/approve access requests to personal/sensitive data)?

Metadata:

Will metadata be made openly available and licenced under a public domain dedication CC0, as per the Grant Agreement? If not, please clarify why. Will metadata contain information to enable the user to access the data?

The metadata will be openly available under the CC BY-SA 4.0 public domain dedication, as this ensures that

credit is given to the metadata creator and that any adaptations of the data are shared under the same terms.

How long will the data remain available and findable? Will metadata be guaranteed to remain available after data is no longer available?

Data will remain available and findable for as long as possible. Metadata will continue to be available even after the data are no longer accessible.

Will documentation or reference about any software be needed to access or read the data be included? Will it be possible to include the relevant software (e.g. in open source code)?

Yes, documentation on the use of INCEpTION for semantic annotation, spaCy for Named Entity Recognition (NER), and digital corpus management tools will be provided in the form of academic papers.

Making data interoperable

What data and metadata vocabularies, standards, formats or methodologies will you follow to make your data interoperable to allow data exchange and re-use within and across disciplines? Will you follow community-endorsed interoperability best practices? Which ones?

Yes. I will follow the best practices of the Linked Open Data (LOD) initiative with a focus on the ancient world.

In case it is unavoidable that you use uncommon or generate project specific ontologies or vocabularies, will you provide mappings to more commonly used ontologies? Will you openly publish the generated ontologies or vocabularies to allow reusing, refining or extending them?

I'm generating project specific vocabularies to annotate authors, works, and verbs of saying related to citations of historians. I will map them to more commonly used ontologies and will openly publish them.

Will your data include qualified references⁴ to other data (e.g. other data from your project, or datasets from previous research)?

Yes, my data include qualified references in the form of RDF triples according to the LOD best practices.

Increase data re-use

How will you provide documentation needed to validate data analysis and facilitate data re-use (e.g. readme files with information on methodology, codebooks, data cleaning, analyses, variable definitions, units of measurement, etc.)?

In the GitHub repository of the project, I will provide a readme file with information on methodology, data cleaning, and analyses. Project guidelines will be published as part of the PhD project and in the form of academic papers.

Will your data be made freely available in the public domain to permit the widest re-use possible? Will your data be licensed using standard reuse licenses, in line with the obligations set out in the Grant Agreement?

Yes, data will be made freely available in the public domain with a CC-BY-SA 4.0 license.

Will the data produced in the project be useable by third parties, in particular after the end of the project?

Yes, data will be reusable by third parties (e.g., scholars, students, publishers).

Will the provenance of the data be thoroughly documented using the appropriate standards?

⁴ A qualified reference is a cross-reference that explains its intent. For example, *X is regulator of Y* is a much more qualified reference than *X is associated with Y*, or *X see also Y*. The goal therefore is to create as many meaningful links as possible between (meta)data resources to enrich the contextual knowledge about the data. (Source: <https://www.go-fair.org/fair-principles/i3-metadata-include-qualified-references-metadata/>)

Yes, data will be documented using the appropriate standards according to the LOD initiative for the Ancient World.

Describe all relevant data quality assurance processes.

Data quality is assured by checking and cleaning data. As these data are mostly in ancient Greek or related to the ancient Greek world, the quality of the data checking and cleaning is assured by my formation as a classical philologist.

Further to the FAIR principles, DMPs should also address research outputs other than data, and should carefully consider aspects related to the allocation of resources, data security and ethical aspects.

Outputs will also include the PhD dissertation and papers in academic journals. Data security will be ensured by storing the data in open and well-established repositories, such as GitHub.

Other research outputs

In addition to the management of data, beneficiaries should also consider and plan for the management of other research outputs that may be generated or re-used throughout their projects. Such outputs can be either digital (e.g. software, workflows, protocols, models, etc.) or physical (e.g. new materials, antibodies, reagents, samples, etc.).

Beneficiaries should consider which of the questions pertaining to FAIR data above, can apply to the management of other research outputs, and should strive to provide sufficient detail on how their research outputs will be managed and shared, or made available for re-use, in line with the FAIR principles.

Allocation of resources

What will the costs be for making data or other research outputs FAIR in your project (e.g. direct and indirect costs related to storage, archiving, re-use, security, etc.)?

So far, there are no costs associated with making the data or other research outputs FAIR in this project.

How will these be covered? Note that costs related to research data/output management are eligible as part of the Horizon Europe grant (if compliant with the Grant Agreement conditions)

Who will be responsible for data management in your project?

I will be responsible for data management.

How will long term preservation be ensured? Discuss the necessary resources to accomplish this (costs and potential value, who decides and how, what data will be kept and for how long)?

Data security

What provisions are or will be in place for data security (including data recovery as well as secure storage/archiving and transfer of sensitive data)?

Will the data be safely stored in trusted repositories for long term preservation and curation?

Ethics

Are there, or could there be, any ethics or legal issues that can have an impact on data sharing? These can also be discussed in the context of the ethics review. If relevant, include references to ethics deliverables and ethics chapter in the Description of the Action (DoA).

Will informed consent for data sharing and long term preservation be included in questionnaires dealing with personal data?

Other issues

Do you, or will you, make use of other national/funder/sectorial/departmental procedures for data management? If yes, which ones (please list and briefly describe them)?

HISTORY OF CHANGES		
VERSION	PUBLICATION DATE	CHANGE
1.0	05.05.2021	Initial version (new MFF).
1.1	01.04.2022	Reformatted to align with other deliverables templates.

6. Recovering anonymous late-antique preachers in the corpus of pseudo- Augustinian sermons: Data Management Plan

Kendall M. Bitner

Radboud University Nijmegen

20 February 2025

This DMP was created in RIS

Data summary

- 1.1 Will you re-use any existing data and what will you re-use it for? State the reasons if re-use of any existing data has been considered but discarded.

The main sources of data will be digitized manuscripts from various archives and libraries and the Patristic Sermons in the Middle Ages (PASSIM) open source research tool (<https://passim.rich.ru.nl>). The manuscripts will be used for textual editing, (automatic) transcription, and research into the transmission patterns of patristic sermons. The PASSIM database will be my main resource for the metadata available for sermons and the manuscripts that contain them.

- 1.2 What types and formats of data will the project generate or re-use?

I expect to generate the following kinds of datasets:

1. Spreadsheets containing various manuscript and text metadata (.csv and/or .xlsx)
2. Manuscript transcriptions. Some of these may be generated by automated transcription software. (.xml and/or other text files)
3. Textual transmission data generated and/or visualized as network maps, graphs, etc. via the PASSIM database (.svg, .png, .json, .xlsx)
4. Text reuse detection analyses, including: a) individual instances of blocks of nearly identical text being reused in various contexts, and b) large-scale patterns of such text reuse within the corpus of patristic sermons. 5. An annotated bibliography on canons, canonicity, and canon formation, managed via the open source reference software, Zotero.

- 1.3 What is the purpose of the data generation or re-use and its relation to the objectives of the project?

To investigate relationships between medieval texts and manuscripts. I expect to uncover many hitherto undetected instances of text reuse within the corpus, which will enrich our understanding of the transmission of texts in the medieval period. I also may find evidence that could challenge conventional applications of canonicity to a medieval context.

- 1.4 What is the expected size of the data that you intend to generate or re-use?

Uncertain, but likely less than 200 GB.

1.5 What is the origin/provenance of the data, either generated or re-used?

The main sources of re-used data will be: the PASSIM database and various libraries and archives of digitized manuscripts

The main sources of generated data will be: PASSIM database, and automatic (AI) transcription software, such as eScriptorium or Transkribus. Zotero will be used to manage the canonicity bibliographical data, which will be populated by me and other members of the MSCA doctoral network.

1.6 To whom might your data be useful ('data utility'), outside your project?

My data may be most useful to those interested in patristic studies, canonization and canon formation, sermon studies, and reception studies. It may also be applicable more broadly to the history of Christianity, manuscript culture, textual transmission, and medieval studies generally.

FAIR data: Making data findable, including provisions for metadata

2.1.1 Will data be identified by a persistent identifier?

Yes. The main source for data generated, the PASSIM database, Handwritten text recognition (HTR) related data will be published via appropriate repositories, such as HTRUnited, Zenodo, and GitHub.

2.1.2 Will rich metadata be provided to allow discovery? What metadata will be created? What disciplinary or general standards will be followed? In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how.

Yes, I will follow the standard practice for sermon and manuscript metadata already employed by PASSIM.

2.1.3 Will search keywords be provided in the metadata to optimize the possibility for discovery and then potential re-use?

Yes. Keywords will either be automatically assigned or manual generated.

2.1.4 Will metadata be offered in such a way that it can be harvested and indexed? Yes. Adhering to discipline standards for text and manuscript metadata will ensure that it can be harvested and indexed.

FAIR data: Making data accessible

2.2.1 Will the data be deposited in a trusted repository?

Yes, in the Radboud Data Repository.

2.2.2 Have you explored appropriate arrangements with the identified repository where your data will be deposited?

Yes, I already have access to the repository for managing my data.

2.2.3 Does the repository ensure that the data is assigned an identifier? Will the repository resolve the identifier to a digital object?

Yes, my data will be Findable via the Radboud Data Repository, which is indexed by search engines on the internet. My data will include rich metadata and have persistent identifiers (DOI).

2.2.4 Will all data be made openly available? If certain datasets cannot be shared (or need to be shared under restricted access conditions), explain why, clearly separating legal and contractual reasons from intentional restrictions. Note that in multi-beneficiary projects it is also possible for specific beneficiaries to keep their data closed if opening their data goes against their legitimate interests or other constraints as per the Grant Agreement.

Yes, all data will be made openly available. No restrictions are applicable.

2.2.5 If an embargo is applied to give time to publish or seek protection of the intellectual property (e.g. patents), specify why and how long this will apply, bearing in mind that research data should be made available as soon as possible. N/A

2.2.6 Will the data be accessible through a free and standardized access protocol?

Yes, managed by the Radboud Data Repository.

2.2.7 If there are restrictions on use, how will access be provided to the data, both during and after the end of the project? N/A

2.2.8 How will the identity of the person accessing the data be ascertained?

N/A

2.2.9 Is there a need for a data access committee (e.g. to evaluate/approve access requests to personal/sensitive data)?

No. No personal or sensitive data is involved.

2.2.10 Will metadata be made openly available and licenced under a public domain dedication CC0, as per the Grant Agreement? If not, please clarify why. Will metadata contain information to enable the user to access the data?

Yes, under a CC-BY license.

2.2.11 How long will the data remain available and findable? Will metadata be guaranteed to remain available after data is no longer available?

I will follow my institute's policy and archive the research data associated with my publication (including raw data, metadata, and documentation) in the Radboud Data Repository for a minimum of 10 years after the end of the project.

2.2.12 Will documentation or reference about any software be needed to access or read the data be included? Will it be possible to include the relevant software (e.g. in open source code)? Yes, for some data generated via PASSIM, e.g. data for network overlap visualizations (.json). However, much of this data will remain reproducible via the PASSIM research tool itself.

FAIR data: Making data interoperable

- 2.3.1 What data and metadata vocabularies, standards, formats or methodologies will you follow to make your data interoperable to allow data exchange and re-use within and across disciplines? Will you follow community-endorsed interoperability best practices? Which ones?

I will follow the standards employed in the PASSIM data model for manuscript and text metadata. Alternatively, I may update the metadata to the most recent ontologies being developed in the field, e.g. by the Manuscript AI research project.

- 2.3.2 In case it is unavoidable that you use uncommon or generate project specific ontologies or vocabularies, will you provide mappings to more commonly used ontologies? Will you openly publish the generated ontologies or vocabularies to allow reusing, refining or extending them?

Data ontologies will follow the latest discipline standards.

- 2.3.3 Will your data include qualified references to other data (e.g. other data from your project, or datasets from previous research)?

Yes, qualified references will be used for data internal to the project and already existing in PASSIM.

FAIR data: Increase data re-use

- 2.4.1 How will you provide documentation needed to validate data analysis and facilitate data re-use (e.g. readme files with information on methodology, codebooks, data cleaning, analyses, variable definitions, units of measurement, etc.)?

Readme files on methodology, data cleaning, and analyses will be provided. Alternatively, already published protocols and discipline standards may be pointed to, e.g. Manuscript AI ontologies and Text Encoding Initiative (TEI) guidelines for transcriptions.

- 2.4.2 Will your data be made freely available in the public domain to permit the widest re-use possible? Will your data be licensed using standard re-use licenses, in line with the obligations set out in the Grant Agreement?

Yes, via a Diamond Open Access license.

- 2.4.3 Will the data produced in the project be useable by third parties, in particular after the end of the project?

Yes. My institute ensures that research data associated with my publication (including raw data, metadata, and documentation) will be archived in the Radboud Data Repository for a minimum of 10 years after the end of the project.

- 2.4.4 Will the provenance of the data be thoroughly documented using the appropriate standards?

Yes, discipline standards will be followed in documenting data provenance.

- 2.4.5 Describe all relevant data quality assurance processes. Data will be curated by the Radboud Data Repository. Data will be scrutinized by experts in the field, both within and beyond my supervisory board.

Other research outputs

- 3.1 In addition to the management of data, beneficiaries should also consider and plan for the management of other research outputs that may be generated or re-used throughout their projects. Such outputs can be either digital (e.g. software, workflows, protocols, models, etc.) or physical (e.g. new materials, antibodies, reagents, samples, etc.). Will your project generate outputs other than data?

Outputs will include: a collective annotated bibliography on canonicity, research articles, blog posts, and handwritten text recognition (HTR) transcriptions and models.

- 3.2 Beneficiaries should consider which of the questions pertaining to FAIR data above, can apply to the management of other research outputs, and should strive to provide sufficient detail on how their research outputs will be managed and shared, or made available for re-use, in line with the FAIR principles. How will you ensure that your research outputs other than data fulfil the FAIR principles?

The annotated bibliography will be contributed to by all members of the MSCA doctoral network and the data will be managed and published via the open source reference software, Zotero. Ideally, the bibliography will also be published in open access by a major scholarly publisher. Research articles will be exclusively published via open access journals or publishers. Blog posts will be published on the MECANO MSCA doctoral network website (<https://mecano-dn.eu/>). HTR models and transcriptions will be published via Zenodo, GitHub, and/or HTRUnited.

Allocation of resources

- 4.1 What will the costs be for making data or other research outputs FAIR in your project (e.g. direct and indirect costs related to storage, archiving, re-use, security, etc.)?

There may be some costs for data storage. Specifically, a paid Zotero subscription may be required for sufficient storage for the annotated bibliography.

- 4.2 How will these be covered? Note that costs related to research data/output management are eligible as part of the Horizon Europe grant (if compliant with the Grant Agreement conditions).

Costs will be covered via Horizon Europe funding.

- 4.3 Who will be responsible for data management in your project? Me, my supervisor, my fellow MSCA doctoral network members (for the collective bibliography), and the Radboud Data Repository.

- 4.4 How will long term preservation be ensured? Discuss the necessary resources to accomplish this (costs and potential value, who decides and how, what data will be kept and for how long)?

I will follow my institute's policy and archive the research data associated with my publication (including raw data, metadata, and documentation) in the Radboud Data Repository for a minimum of 10 years after the end of the project.

Data security

- 5.1 What provisions are or will be in place for data security (including data recovery as well as secure storage/archiving and transfer of sensitive data)?

No sensitive data is involved. The Radboud Data Repository and the PASSIM database will manage data storage, archiving, and recovery.

5.2 Will the data be safely stored in trusted repositories for long term preservation and curation?

Yes, in the Radboud Data Repository for a minimum of 10 years after the end of the project. Much data will be additionally stored in the PASSIM database.

Ethics

- 6.1 Are there, or could there be, any ethics or legal issues that can have an impact on data sharing? These can also be discussed in the context of the ethics review. If relevant, include references to ethics deliverables and ethics chapter in the Description of the Action (DoA).

No

- 6.2 Will informed consent for data sharing and long term preservation be included in questionnaires dealing with personal data?

N/A

Other issues

- 7.1 Do you, or will you, make use of other national/funder/sectorial/departmental procedures for data management? If yes, which ones?

I will follow both the Horizon Europe and Radboud University guidelines for procedures on data management.

7. Syntax, formulaic structures, and canon-marking in Greek and Arabic: documentary texts and Galen

DATA MANAGEMENT PLAN

PROJECT	
Project number:	[project number]
Project acronym:	[acronym]
Project name:	MECANO and Lexical Evolution: Tracing Semantic Shifts from Ancient Greek to Arabic in Medical Texts

DATA MANAGEMENT PLAN	
Date:	20/02/2025
Version:	1.0

Data Summary

This project will re-use historical and linguistic data from select passages from Greek pharmacological and medical texts (papyrus, manuscript, and editions), as well as Arabic translations and a compendium, in both unedited manuscripts and in editions, with the aim of analyzing semantic and structural shifts in medical terminology and formulae. Some later indirect citations and references of these sources, as well as comparanda, will be analyzed from medieval medical authors writing in Arabic. Occasional reference may possibly be made to Arabic-into-Latin translation. These sources form the foundation for tracing the evolution of medical knowledge across cultural, linguistic,

geographical and diachronic boundaries.

The project will, for instance, examine the manuscripts of Greek original and Arabic translations of:

- **Dioscorides' *De Materia Medica* (1st century CE):** This foundational Greek work systematically catalogues medicinal plants, substances, and their uses. It is a key source for understanding how

medical knowledge was conceptualized in ancient Greek society, which would later be translated and adapted in the Arabic-speaking world.

- **Galen's *De Simplicium Medicamentorum Temperamentis Ad Facultatibus* (2nd century CE):** Galen's medical writings expanded on Hippocratic thought, focusing on the classification of drugs based on their temperaments and their therapeutic potential. His influence persisted in both Greco-Roman and Islamic medical traditions, and his works were extensively translated into Arabic, often with significant adaptations.

as well as a source from the Arabic exegetical and summary literature, namely:

- **Ibn al-Bayṭar's *Kitāb al-Ğāmi' fī al-Adwiya al-Mufrada* (13th century CE):** A detailed compendium of medicinal plants, Ibn al-Bayṭar's work builds on Dioscorides and Galen while integrating his own observations. His systematic classification of medicinal substances is one of the most important contributions to pharmacology in the Islamic world, influencing later European herbalists.

as well as editions (and where needed manuscripts) of the following original Arabic medical works, which refer obliquely or in a derivative manner to the Greek sources, and may possibly reflect acquaintance with the Arabic translations of Dioscorides, Galen and other Greek medical and pharmacological knowledge to which they refer, sometimes with attribution:

- **Ibn Sīnā's *Al-Qānūn fī al-Ṭibb* (The Canon of Medicine, 11th century CE):** A monumental text that synthesized Greek, Roman, Persian, and Arabic medical knowledge, Ibn Sina's work became the standard reference for medical students and scholars in both the Islamic world and Medieval Europe. His refinements to Galenic concepts and his development of new terminologies were essential to the medical lexicon in the Arabic-speaking world.
- **Al-Razi's *Kitāb al-Hawī* (The Comprehensive Book of Medicine, 10th century CE):** This Persian physician's encyclopedic work, written in Arabic, draws heavily on Greek and Roman traditions but is unique in its incorporation of empirical observations and experimental pharmacology. It represents a key stage in the transmission and transformation of pharmacological knowledge, especially in terms of categorizing substances and their effects.

Possible occasional control checks from Arabic-into-Latin translation (editions where available, and if not, manuscripts) will be made by consultation ad locum to terms in

- **Gerard of Cremona's 12th c. CE Latin translations of the medieval Arabic versions of Galen's *De Simplicium Medicamentorum Temperamentis Ad Facultatibus***

This project will analyze how these medical texts were translated and interpreted across cultures, exploring the lexical shifts and semantic transformations that occurred in the process. By examining the terminology and classification systems employed in these texts, the project aims to shed light on the evolution of pharmacological thought and practice, revealing the **interactions between different medical traditions** and their long-term impact on global medical knowledge.

FAIR data

2.1. Making data findable, including provisions for metadata 2.1. Making Data Findable, Including Provisions for Metadata

- **Persistent Identifiers (PIDs):**

All the data from this project will have unique **identifiers**, like **DOIs (Digital Object Identifiers)**, so that people can easily find and reference them in the future.

- **Metadata to Help Discover the Data:**

We will create detailed **metadata** (information about the data) to make them easy to find. The metadata will include: o **Basic Information:** Like the title, author(s), date, language, and what the text is about.

- o **Contextual Information:** Details about the historical and linguistic background of the texts (e.g., Greek and Arabic medical traditions).
- o **Key Terms:** Information about important medical terms, especially how they change in meaning over time and across languages.
- o **Translations and Annotations:** Notes on translations and explanations of medical terms.
- **Metadata Standards:**

We will follow established **metadata standards** to make sure the data is easy to use. This includes common systems like **Dublin Core** (for basic details) and **TEI (Text Encoding Initiative)**, which is often used for ancient texts.

If no existing standards cover our specific needs, we will create a **custom metadata schema** to track things like changes in medical terminology and other project-specific details.

- **Keywords to Help with Discovery:**

We will include important **keywords** in the metadata (such as **Greek pharmacology**, **semantic shifts**, **Arabic medical texts**, etc.) to make sure that people searching for related topics can easily find our data.

- **Harvesting and Indexing the Metadata:**

We will make the metadata available in a format that can be collected by other research platforms (like **Zenodo**, **OpenAIRE**, etc.), so they can be easily discovered by researchers around the world.

Repository:

- The data will be deposited in a trusted repository to ensure long-term accessibility and preservation.
- We have explored appropriate arrangements with the identified repository where the data will be deposited. The repository we plan to use is **Zenodo**, which provides persistent identifiers for datasets and ensures they are resolvable to digital objects. However, we will also explore other suitable repositories like **OpenAIRE**, **DANS**, or **CLARIN** depending on the specific needs of the project.

Data:

- All data will be made openly available, with proper metadata provided for discovery and re-use.
- Some data, especially if it involves sensitive or unpublished medical data, might be subject to embargo periods to protect intellectual property or pending publications. In such cases, access will be granted under restricted conditions, and the data will be made available as soon as possible, post-embargo.
- The data will be accessible through a free and standardized access protocol such as OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting).

Metadata:

- Metadata will be made openly available and licensed under a public domain dedication (CC0), in compliance with the Grant Agreement. The metadata will contain sufficient information to allow users to access the data, including the relevant DOI and access links.

- The metadata and data will remain available and findable for as long as the repository's standards allow, ensuring that the data remains accessible to users in the long term.
- Documentation or reference to any software needed to access or read the data will be included, and open-source code for the relevant software will be provided where applicable.

Making data interoperable

Data and Metadata Standards:

- To ensure the data is easy to share and understand, we will use widely accepted standards and formats:
 - **TEI (Text Encoding Initiative)** for encoding Greek and Arabic medical texts.
 - **CSV** and **SQL** for databases of medical terms and their meanings.
- **JSON** and **XML** for storing and sharing metadata.
- We will follow **FAIR principles** (Findable, Accessible, Interoperable, Reusable) to make our data easy to find and use.

Use of Standardized Vocabularies:

- We will use established medical vocabularies like:
 - **MeSH (Medical Subject Headings)** to standardize medical terms.
 - **SKOS (Simple Knowledge Organization System)** for showing relationships between terms.
- If we create any new terms specific to our project, we will map them to these recognized vocabularies and make them publicly available.

References to Other Data:

- We will link our data to other relevant datasets, whether from our project or external sources, to provide further context and connections to other research.

CLARIN for Interoperability:

- Given the linguistic nature of the project, **CLARIN** will be used for resources and tools that support linguistic research and ensure our data is interoperable across different disciplines.

Increase data re-use

To ensure the data generated in this project can be validated, reused, and contribute to broader academic research, the following documentation and practices will be implemented:

- **Documentation for data validation and reuse:** The project will provide detailed documentation, including readme files, metadata, and codebooks, which will describe the methodology, transcription processes, data cleaning procedures, and annotations. This will allow future users to understand the context, methods, and structure of the data, ensuring their validity and usability.
- **Open data availability:** The data, including transcriptions of ancient Greek and Arabic medical texts, lexical databases, and annotated glossaries, will be made freely available to the public, with a license that encourages academic reuse. The project will use open-access platforms like Zenodo, ensuring that the data are shared under licenses that align with the Grant Agreement, such as CC BY or CC0.

- **Usability for third parties:** The datasets generated will be reusable by third parties, particularly scholars and researchers in fields such as history of medicine, linguistics, and digital humanities. The data will remain accessible after the end of the project, allowing continued use and analysis.
- **Data provenance:** The project will document the origin of the data sources meticulously, ensuring that the provenance of each dataset is clear. This will include references to original texts and translations, such as those by Galen, Dioscorides, Ibn Sina, and others, as well as modern lexicons and databases used for reference.
- **Data quality assurance:** To ensure high-quality data, the project will implement a rigorous data cleaning process. This includes verifying the accuracy of the transcriptions, ensuring proper encoding of metadata, and using reliable sources for creating lexical databases. The project will follow established practices for data quality assurance in digital humanities and historical linguistics.

Other research outputs

In addition to managing the main project data, the project will produce *Derivatio Canonica Lexici* (The Canonical Derivation of the Lexicon), a digital lexicon tracing the evolution of medical and pharmacological terminology between Greek and Arabic. It will focus on key figures such as Galen, Dioscorides, Ibn Sīnā, and Rāzī.

Key Features:

- **Digital Access:** The lexicon will be available in open, structured formats for easy access and use.
- **Cross-Referenced:** It will link with other datasets and research to support interdisciplinary studies.
- **Clear Documentation:** Instructions on methodology and usage will ensure others can understand and reuse the lexicon.

FAIR Principles:

- **Findable:** The lexicon will have a DOI for easy citation and discovery.
- **Accessible:** It will be deposited in an open-access repository for free global access.
- **Interoperable:** It will follow standard metadata formats for easy integration with other research outputs.
- **Reusable:** Comprehensive documentation will ensure the lexicon can be reused in future research.

This lexicon will serve as an important tool for scholars studying medical history, language evolution, and intercultural knowledge exchange.

Allocation of resources

To make the data and research outputs of this project FAIR, several resource needs will arise, including:

- **Direct Costs:** This will include expenses for storing and archiving data, ensuring data security, and enabling long-term access. Costs will also cover maintaining databases, ensuring proper metadata documentation, and ensuring the usability of the data for future research.
- **Indirect Costs:** These may involve administrative support, the integration of collaborative efforts, and additional research outreach, particularly with academic institutions, to ensure the sustainability and re-use of the project's outputs.

These costs will be covered by the **Horizon Europe grant** and supplementary funding from **academic collaborations** with institutions in the Gulf region. Institutions like **Qatar University**, **King Saud University**, and **The American University of Sharjah** may provide both financial and logistical support, ensuring the preservation of the research outputs.

Responsibility for Data Management: The responsibility for data management will be handled by the project team, with a designated **Data Manager** overseeing all aspects of data handling, from collection to preservation. This individual will ensure that all data adheres to the FAIR principles and that processes for archiving and sharing are transparent and effective.

Long-Term Preservation: Long-term preservation will be ensured by depositing the project data in trusted repositories such as **Zenodo** and **CLARIN**, both of which are known for maintaining open access and long-term accessibility. In addition to digital archiving, academic partnerships in the Gulf region will help support continuous data accessibility.

The data will be preserved for a minimum of **10 years** to ensure its long-term relevance. Decisions regarding data retention will be made in collaboration with the institutions involved in the project, ensuring that only the most relevant datasets (such as transcribed texts, lexical databases, and annotated translations) are maintained. These datasets will remain open to the academic community for future re-use and analysis.


Data security

- **Data Security Measures:** All project data will be securely stored using trusted repositories like **Zenodo** and **CLARIN**, with regular backups to ensure data recovery in case of failure.
- **Sensitive Data:** Since the project deals with historical texts and translations, there are no concerns about personal or highly sensitive data. However, all data will be handled carefully and securely.
- **Long-Term Preservation:** Data will be archived in these repositories for long-term preservation and future access. Access will be controlled based on the project's needs, ensuring the data remains safe and retrievable.

Ethics

- **Ethical Considerations:** This project deals exclusively with historical and linguistic data from ancient Greek and Arabic medical texts, all of which are publicly available or in the public domain. Since no personal data or human subjects are involved, there are no direct ethical concerns regarding privacy or informed consent. Additionally, the research is secular in nature and will be conducted with a strict commitment to academic objectivity and impartiality. The project will show no bias towards any particular culture, religion, or historical perspective.
- **Cultural Sensitivity:** While there is no personal data in this project, the materials we work with might have cultural or historical significance. We will ensure that our work respects the intellectual and cultural heritage of the texts and any modern descendants or communities associated with them. Special care will be taken to avoid misinterpretation or misrepresentation of historical medical practices, maintaining neutrality and respect throughout the research.
- **Data Sharing & Informed Consent:** Although this project does not involve human data, if any future work requires access to sensitive material (e.g., unpublished manuscripts or new findings), informed consent or proper permissions will be sought in line with ethical best practices.
- **Ethics Review:** Given the nature of the project, we do not foresee significant ethical concerns, but we will conduct a review of relevant academic, institutional, and legal standards regarding the use and sharing of ancient texts and translations. This will ensure that our methodology is transparent, ethical, and compliant with the relevant laws on intellectual property.


- **Data Use and Licensing:** Data shared through repositories like Zenodo or CLARIN will be openly available under **open access** or **CC0 licenses**, in line with the principles of openness, transparency, and reusability. Proper attribution will always be provided to the original sources.
- **Respect for Intellectual Property:** Since the data used in this project is either in the public domain or properly licensed, we will ensure that all original sources, translations, and associated materials are properly attributed. We will also respect copyright laws and avoid using restricted content unless appropriate permissions are obtained.

HISTORY OF CHANGES		
	PUBLICATION DATE	CHANGE
1.0	20.02.2025	Initial version (new MFF). Initial version created for the Data Management Plan (DMP) 
1.1	[Date of next update]	Reformatted to align with other deliverables templates.

8. Ancient sources on matter in Late Medieval Commentaries on Aristotle

DATA MANAGEMENT PLAN

(To be filled in and uploaded as deliverable in the Portal Grant Management System, at the due date foreseen in the system (and regularly updated).)

 *The template is recommended but not mandatory. If you do not use it, please make however sure that you comply with the research data management requirements under Article 17 of the Grant Agreement.)*

PROJECT	
Project number:	[101120349]

Project acronym:	[MECANO]
Project name:	[The Mechanics of Canon Formation and the Transmission of Knowledge from Graeco-Roman Antiquity]

DATA MANAGEMENT PLAN	
Date:	[15/02/2025]
Version:	[DMP version 1]

Data Summary

Will you re-use any existing data and what will you re-use it for? State the reasons if re-use of any existing data has been considered but discarded.

Yes, the project will primarily rely on existing data, including both printed books and digital databases. The key sources include:

- **A wide variety of printed books and articles** – These will serve as the primary sources for the project, forming the backbone of the analysis of medieval commentaries on Aristotle's predecessors.
- **Brepolis (Library of Latin Texts & Aristoteles Latinus series)** – for accessing digitized editions of medieval Latin texts.
- **Corpus Corporum database** – This resource will be used for exploring the Latin texts of various authors, providing additional insight into the language and context of the medieval commentaries.
- **Albertus Magnus Search Tool (University of Waterloo)** – for searching Albert the Great's corpus efficiently.
- **Thesaurus Linguae Graecae (TLG)** – for consulting Aristotle's Greek texts and their intertextual references.
- **LOGEION** – for accessing Latin dictionaries relevant to medieval Latin vocabulary and textual interpretation.
- **Index Thomisticus** – This index will be used to access the complete works of Thomas Aquinas, providing critical references for understanding his interpretations of Aristotle and his predecessors.

These resources will be used for analyzing how late medieval commentators interpreted Aristotle's references to his predecessors in Metaphysics and Physics, particularly concerning medieval theories of matter. The re-use of these sources is necessary because they provide authoritative, well-curated editions of relevant texts.

Note: As the project progresses, additional resources may be added depending on the needs of the research, as the project is still in its early stages.

What types and formats of data will the project generate or re-use?

- **Re-used data:** Digital and print texts in Latin and Greek from the above-mentioned databases. These are primarily in TEI-XML, PDF, and TXT formats, depending on the database.
- **Generated data:**
 - 1) Annotated textual analyses in Word (DOCX), plain text (TXT), or LaTeX format.
 - 2) Personal notes, commentary, and transcriptions from medieval Latin manuscripts (if needed), stored

as DOCX, TXT, or Markdown (MD).

3) Bibliographical data, likely managed in reference management software (e.g., Zotero, stored in RIS or BibTeX format).

What is the purpose of the data generation or re-use and its relation to the objectives of the project?

The data is essential for examining how medieval commentators interpreted Aristotle's references to earlier philosophers. By systematically analyzing texts from these sources, the project will:

- Trace conceptual shifts in the understanding of Aristotle's predecessors.
- Identify implicit sources and influences in medieval commentaries.
- Explore how medieval theories of matter shaped these interpretations.
- Produce a structured analysis of key passages in Albert the Great, Thomas Aquinas, Roger Bacon and other relevant commentators.

What is the expected size of the data that you intend to generate or re-use?

- Re-used data: The databases contain several gigabytes of text data, but only specific excerpts relevant to the project will be stored or cited.
- Generated data: The project will produce hundreds of pages of textual analysis, annotations, and references, amounting to approximately 100–500 MB.

What is the origin/provenance of the data, either generated or re-used?

- Re-used data comes from Brepolis, TLG, LOGEION, Index Thomisticus, the Albertus Magnus search tool (with the possibility of including more digital databases) which provide well-documented and authoritative editions of texts, as well as various other functionalities (for example, keyword search tools, the option to find comparable sentences in several different authors, etc.)
- Generated data originates from the researcher's personal analysis and annotations of these texts.

To whom might your data be useful ('data utility'), outside your project?

- Scholars in medieval philosophy and classical reception studies, particularly those studying Aristotelianism and its medieval interpretations.
- Historians of medieval science and metaphysics, who examine theories of matter.
- Philologists and textual scholars, working on Latin and Greek manuscript traditions.
- Researchers in digital humanities, interested in computational approaches to medieval philosophical texts.

FAIR data

Making data findable, including provisions for metadata

Will data be identified by a persistent identifier?

Yes, where possible, data will be assigned persistent identifiers (PIDs) such as DOI (Digital Object Identifier) for published works or other relevant identifiers for key digital resources. This will ensure that datasets and texts used in the project are traceable and can be reliably cited.

Will rich metadata be provided to allow discovery? What metadata will be created? What disciplinary or general standards will be followed? In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how.

Yes, rich metadata will be provided to enable discovery. Metadata will include:

- Title of the resource (e.g., text or manuscript)

- Author(s)
- Edition/version of the text used
- Date of publication or manuscript
- Keywords related to themes (e.g., Aristotelian metaphysics, medieval commentary, matter theory, etc.)
- Abstracts and brief descriptions of each resource's relevance to the research
- Citation information for proper attribution of texts and articles.

Will search keywords be provided in the metadata to optimize the possibility for discovery and then potential re-use?

Yes, relevant search keywords will be included in the metadata to optimize discoverability. These keywords will cover philosophical concepts, key figures (e.g., Aristotle, Albert the Great, Thomas Aquinas), and themes related to the research (e.g., metaphysics, theories of matter, Aristotelian predecessors).

Will metadata be offered in such a way that it can be harvested and indexed?

Yes, metadata will be offered in machine-readable formats (e.g., XML, JSON, or RDF) to ensure it can be harvested and indexed by relevant databases, search engines, and repositories. This will allow other researchers to find and potentially reuse the data in the future.

Making data accessible

Repository:

Will the data be deposited in a trusted repository?

Yes, the data will be deposited in trusted repositories, where it will be securely stored and accessible to others. Potential repositories include:

- Zenodo, a widely used open-access repository that supports a wide variety of data types, including text-based research outputs.
- Dataverse or other institutional repositories at the University of Helsinki and KU Leuven may also be considered for academic content.

Have you explored appropriate arrangements with the identified repository where your data will be deposited?

Yes, initial exploration has been done, and appropriate arrangements will be made with the identified repository (such as Zenodo or an institutional repository). The specific choice of repository will depend on the data type and the repository's suitability for long-term storage, accessibility, and the specific academic standards adhered to by the project.

Does the repository ensure that the data is assigned an identifier? Will the repository resolve the identifier to a digital object?

Yes, the chosen repository will ensure that the data is assigned a persistent identifier, such as a DOI (Digital Object Identifier), which will allow the data to be reliably cited and referenced. The repository will resolve the identifier to a digital object, making it easily accessible to other researchers and ensuring long-term findability.

Data:

Will all data be made openly available? If certain datasets cannot be shared (or need to be shared under restricted access conditions), explain why, clearly separating legal and contractual reasons from intentional restrictions. Note that in multi-beneficiary projects it is also possible for specific beneficiaries to keep their data closed if opening their data goes against their legitimate interests or other constraints as per the Grant Agreement.

Yes, the data will be made openly available, subject to any legal, ethical, or contractual restrictions. If certain datasets cannot be shared openly, the reasons for this will be clearly explained, distinguishing between legal and contractual reasons and intentional restrictions. For example, if any data is subject to intellectual property protection or confidentiality agreements, it will be noted.

If an embargo is applied to give time to publish or seek protection of the intellectual property (e.g. patents), specify why and how long this will apply, bearing in mind that research data should be made available as soon as possible.

If an embargo is applied, it will be to allow sufficient time for the publication of results or to seek protection of intellectual property, such as through patents. The embargo period will be limited to the minimum necessary time to achieve these goals, and the data will be made available as soon as possible after that period. The specific duration of the embargo will depend on the publication or intellectual property protection process.

Will the data be accessible through a free and standardized access protocol?

Yes, the data will be accessible through free and standardized access protocols, such as HTTP or OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting). This will ensure that other researchers can easily access and reuse the data.

If there are restrictions on use, how will access be provided to the data, both during and after the end of the project?

If there are any restrictions on the use of data (e.g., due to intellectual property or ethical concerns), access will be controlled through repository settings, ensuring that only authorized users can access restricted datasets. During the project, access may be provided to specific individuals or groups based on their need to know. After the end of

the project, these restrictions will be clearly indicated, and access will be provided in accordance with the project's access policy.

How will the identity of the person accessing the data be ascertained?

The identity of those accessing restricted data will be ascertained through the authentication mechanisms provided by the chosen repository, such as using institutional logins or personal accounts linked to recognized academic institutions.

Is there a need for a data access committee (e.g. to evaluate/approve access requests to personal/sensitive data)?

Given the nature of this project, which does not involve personal or sensitive data, there is no need for a data access committee. However, if any sensitive data or restricted datasets are involved later in the project, a committee or access evaluation process may be established.

Metadata:

Will metadata be made openly available and licenced under a public domain dedication CC0, as per the Grant Agreement? If not, please clarify why. Will metadata contain information to enable the user to access the data?

- Yes, the metadata will be made openly available and licensed under the CC0 public domain dedication, ensuring that the metadata can be freely used by anyone without any restrictions. If any exceptions are made, the reasons will be explained.
- Yes, metadata will include clear instructions on how to access the relevant data, including persistent identifiers (such as DOIs) and direct links to datasets or documents.

How long will the data remain available and findable? Will metadata be guaranteed to remain available after data is no longer available?

The data will remain available for as long as possible, ideally indefinitely, to ensure that it can be accessed by future researchers. The metadata will remain available even after the data is no longer accessible, so that users can track and cite the data even after its removal or change in status.

Will documentation or reference about any software be needed to access or read the data be included? Will it be possible to include the relevant software (e.g. in open source code)?

Yes, any necessary software or tools to access or interpret the data will be documented. If the software is available in open-source format, it will be included in the repository along with the data. If specific tools are required for analysis (such as for text encoding or analysis), relevant documentation and instructions for use will be provided.

Making data interoperable

What data and metadata vocabularies, standards, formats or methodologies will you follow to make your data interoperable to allow data exchange and re-use within and across disciplines? Will you follow community-endorsed interoperability best practices? Which ones?

To ensure that the data is interoperable and can be exchanged and reused across disciplines, the project will adhere to well-established metadata standards such as Dublin Core for bibliographic data and Text Encoding Initiative (TEI) for encoding texts. The data will be provided in standard, machine-readable formats like XML, JSON, or CSV, depending on the specific data type. We will also follow best practices for interoperability endorsed by the OpenAIRE guidelines and the FAIR principles to ensure that the data can be easily integrated into various repositories and databases.

Additionally, we will follow community-endorsed best practices for digital humanities, such as ensuring that the data is structured in a way that allows easy reuse across various platforms. The project will also follow the relevant FAIR data principles to enhance the discoverability and usability of the data for other researchers.

In case it is unavoidable that you use uncommon or generate project specific ontologies or vocabularies, will you provide mappings to more commonly used ontologies? Will you openly publish the generated ontologies or vocabularies to allow reusing, refining or extending them?

If project-specific ontologies or vocabularies are created, they will be mapped to widely used ontologies where possible. For example, Library of Congress Subject Headings (LCSH) or other relevant controlled vocabularies will be used where appropriate. We will also make any generated ontologies or vocabularies openly available, either by publishing them in the project's repository or by linking to external platforms where they can be reused, refined, and extended by other researchers in the field.

Will your data include qualified references⁵ to other data (e.g. other data from your project, or datasets from previous research)?

Yes, the data will include qualified references to other relevant data, both from within the project and from previous research. These references will be included as part of the metadata, ensuring that the data is connected to relevant prior work and that the relationships between datasets are clearly documented. This will allow for the re-use and further exploration of related datasets.

⁵ A qualified reference is a cross-reference that explains its intent. For example, *X is regulator of Y* is a much more qualified reference than *X is associated with Y*, or *X see also Y*. The goal therefore is to create as many meaningful links as possible between (meta)data resources to enrich the contextual knowledge about the data. (Source: <https://www.go-fair.org/fair-principles/i3-metadata-include-qualified-references-metadata/>)

Increase data re-use

How will you provide documentation needed to validate data analysis and facilitate data re-use (e.g. readme files with information on methodology, codebooks, data cleaning, analyses, variable definitions, units of measurement, etc.)?

Comprehensive documentation will be provided alongside the data to ensure that others can validate the analysis and reuse the data effectively. This documentation will include readme files that detail the project's methodology, including how the texts are analyzed, how sources are interpreted, and any key assumptions. This will help facilitate understanding and re-use of the data.

Will your data be made freely available in the public domain to permit the widest re-use possible? Will your data be licensed using standard reuse licenses, in line with the obligations set out in the Grant Agreement?

Yes, the data will be made freely available in the public domain, with licensing under an open Creative Commons (CC0) license, which will permit the widest possible re-use. This aligns with the requirements set out in the Grant Agreement and ensures that the data can be reused, adapted, and shared without restrictions. If a different license is required by the grant, this will be specified in the documentation.

Will the data produced in the project be useable by third parties, in particular after the end of the project?

Yes, the data will be designed and formatted for easy usability by third parties both during and after the project. The documentation will ensure that the dataset is well-structured and clearly described, allowing external researchers to understand and use the data effectively even after the project concludes.

Will the provenance of the data be thoroughly documented using the appropriate standards?

Yes, the provenance of the data will be thoroughly documented, with clear records of its creation, transformation, and any external sources used. This will be done using standard metadata formats and will include references to the original sources of the data, the steps taken to process it, and the contributors involved. This documentation will enable others to trace the data's origins and assess its quality.

Describe all relevant data quality assurance processes.

Data quality assurance will be maintained through several processes:

- Validation of sources: The data will be sourced from trusted databases and academic sources, ensuring that the original texts and translations are of high quality.
- Consistency checks: The analysis of the data will involve consistency checks to ensure that interpretations and references are accurate and reliable.
- Peer review: Key datasets and the methodology will be peer-reviewed as part of the academic process to ensure reliability and validity.
- Version control: Data versions will be carefully tracked, and any updates or changes will be recorded to maintain the integrity of the dataset.

Further to the FAIR principles, DMPs should also address research outputs other than data, and should carefully consider aspects related to the allocation of resources, data security and ethical aspects.

In addition to data, the project will consider the broader scope of research outputs, including publications, presentations, and software tools developed as part of the project. These will also be made publicly available, with appropriate licenses.

- Allocation of resources: Adequate resources will be allocated for the documentation and storage of the data, and these will be included in the project budget.
- Data security: Data will be securely stored in trusted repositories with encrypted backups, ensuring that it is protected from loss or unauthorized access.
- Ethical aspects: Ethical considerations have been taken into account to ensure that the data does not include personal or sensitive information, and any relevant ethical approval will be obtained for any data collection that

requires it.

Other research outputs

In addition to the management of data, beneficiaries should also consider and plan for the management of other research outputs that may be generated or re-used throughout their projects. Such outputs can be either digital (e.g. software, workflows, protocols, models, etc.) or physical (e.g. new materials, antibodies, reagents, samples, etc.).

In addition to data, this project will generate several other types of research outputs, primarily in digital formats, including scholarly publications, presentations, and potentially some workflows related to the textual analysis of medieval commentaries. The project will not generate any physical outputs, such as materials, reagents, or samples. All research outputs will be carefully managed and shared following the FAIR principles to ensure that they are reusable by others, both during and after the project.

- **Publications:** Research articles, books, and conference papers will be the primary digital outputs. These will be made available via open-access repositories and journals where applicable. Full metadata (e.g., DOI, authorship, keywords) will be provided to make these outputs findable.
- **Presentations:** Slides, posters, and other materials from presentations will be shared through institutional repositories or personal websites when possible, along with appropriate metadata for easy discoverability.
- **Digital Tools/Workflows:** If any digital tools (such as scripts for textual analysis) are developed, they will be shared openly under an open-source license (e.g., MIT or GPL), and hosted on platforms like GitHub, accompanied by comprehensive documentation.

Beneficiaries should consider which of the questions pertaining to FAIR data above, can apply to the management of other research outputs, and should strive to provide sufficient detail on how their research outputs will be managed and shared, or made available for re-use, in line with the FAIR principles.

The following steps will be taken to ensure that the other research outputs generated during the project are managed and shared according to the FAIR principles:

- **Findability:** All research outputs will be assigned persistent identifiers (e.g., DOI for publications, URN for presentations), ensuring they are easily discoverable. Metadata (title, authors, keywords, etc.) will be used to describe these outputs, and proper indexing will be arranged to ensure that the materials can be easily found.
- **Accessibility:** All research outputs will be openly available and accessible through standardized access protocols, with no restrictions unless required for legal or ethical reasons. Open access repositories and platforms like GitHub will be used to host materials.
- **Interoperability:** Research outputs, especially digital tools, will be designed using standard formats that allow for interoperability. If software or workflows are developed, they will follow community standards for code and documentation, ensuring they are compatible with other research tools.
- **Reusability:** All outputs will be licensed for reuse (e.g., CC0 for publications, open-source licenses for tools), and documentation will ensure that others can reuse and adapt the research materials. Detailed readme files and usage instructions will be included for any digital tools or workflows, ensuring they are easily understandable and reusable.

Allocation of resources

What will the costs be for making data or other research outputs FAIR in your project (e.g. direct and indirect costs related to storage, archiving, re-use, security, etc.) ?

The costs for ensuring that the data and other research outputs are FAIR will be relatively low, as the project primarily relies on printed books and existing digital databases. However, there will be some expenses related to:

- **Data storage:** Hosting and archiving research outputs (publications, presentations, tools) on open-access platforms and institutional repositories. These costs will primarily involve the use of institutional infrastructure and external repositories, which might have a small fee for storage or access (e.g., on GitHub or Zenodo).
- **Metadata creation:** The creation of metadata for publications and digital tools will require some time investment to ensure proper description and indexing, but it will not involve significant direct financial costs.
- **Software and workflow development:** If any software or workflows are developed during the project, costs for hosting and maintaining them (e.g., GitHub) may incur, although the open-source nature of the project will keep these costs minimal.
- **Security and backups:** Ensuring that all digital outputs are securely stored, including regular backups of project data (where applicable), will require minimal resources, mainly provided by the hosting institutions.

How will these be covered? Note that costs related to research data/output management are eligible as part of the Horizon Europe grant (if compliant with the Grant Agreement conditions).

These costs will be covered by the Horizon Europe grant, as they fall within the eligible categories for research data/output management. The budget for these expenses will be allocated in the project's financial plan under the relevant cost categories, ensuring compliance with the Grant Agreement.

Who will be responsible for data management in your project?

The responsibility for data management in this project will be shared between myself and the project supervisors at University of Helsinki and KU Leuven. I will take the lead in ensuring the proper documentation of research outputs, while the supervisors will provide guidance on how best to manage, store, and share these materials in line with FAIR principles.

How will long term preservation be ensured? Discuss the necessary resources to accomplish this (costs and potential value, who decides and how, what data will be kept and for how long)?

Long-term preservation of data and research outputs will be ensured through the use of trusted repositories such as Zenodo or institutional repositories, both of which are designed to ensure long-term accessibility. The necessary resources for this include:

- **Costs:** There will be minimal direct costs for long-term preservation, as the project will use open-access repositories and institutional infrastructure. However, any additional costs related to long-term storage (e.g., for larger digital files) will be covered by the Horizon Europe funding.
- **Value:** The long-term preservation of research outputs will ensure that they remain accessible and reusable by future researchers. This is crucial for ensuring the impact and continuation of the research, particularly as it pertains to the analysis of medieval commentaries on Aristotle, which may continue to be a source of academic interest.
- **Decision-Making:** Decisions regarding long-term preservation will be made in consultation with the project supervisors and the University of Helsinki and KU Leuven digital archiving teams, who will provide guidance on the best practices for ensuring data sustainability and availability.

The research outputs will be kept for at least 10 years after the completion of the project, as per Horizon Europe guidelines. This timeframe will ensure that the data and publications remain available for further academic use and re-analysis.

Data security

What provisions are or will be in place for data security (including data recovery as well as secure storage/archiving and transfer of sensitive data)?

The security of the research data will be ensured through a combination of secure storage, encryption, and backup procedures. The main provisions for data security include:

- **Data Recovery:** Regular backups of research data (both digital textual analyses and metadata) will be conducted to prevent data loss. These backups will be stored in secure environments, both locally (on institutional servers) and remotely (cloud storage services), to ensure recovery in case of accidental loss or technical issues.
- **Secure Storage and Archiving:** The data will be stored in trusted repositories, such as Zenodo, and institutional repositories at University of Helsinki and KU Leuven, which are designed to ensure data integrity, security, and compliance with long-term preservation standards. These repositories have established security measures including encryption, access control, and regular system audits.
- **Transfer of Sensitive Data:** Since no sensitive personal data is involved in the project, there is no immediate risk related to transferring sensitive data. For any data transfers that might occur, secure file transfer protocols (e.g., SFTP, HTTPS) will be used to ensure data security during the transfer process. All data will be anonymized or devoid of personally identifiable information.

Will the data be safely stored in trusted repositories for long term preservation and curation?

Yes, the data will be safely stored in trusted repositories for long-term preservation and curation. Both Zenodo and institutional repositories at University of Helsinki and KU Leuven comply with high standards for digital curation and long-term preservation. They ensure that data remains accessible and usable for the long term, even after the completion of the project. The data will be indexed with proper metadata and assigned persistent identifiers (e.g., DOIs), ensuring it remains available for future re-use and academic citation.

Ethics

Are there, or could there be, any ethics or legal issues that can have an impact on data sharing? These can also be discussed in the context of the ethics review. If relevant, include references to ethics deliverables and ethics chapter in the Description of the Action (DoA).

Given that this project focuses on the analysis of historical texts and does not involve personal data, there are no immediate ethical or legal concerns that would affect data sharing. The project will utilize publicly available texts and digital resources, which are either in the public domain or accessible through repositories with appropriate licensing. As no sensitive personal data is being collected or processed, there are no ethics-related issues to address in this regard.

In the case that any unforeseen issues arise (e.g., if unpublished materials are used), the relevant ethics review processes at University of Helsinki and KU Leuven will be followed to ensure compliance with institutional and national standards.

Will informed consent for data sharing and long term preservation be included in questionnaires dealing with personal data?

Since no personal data is involved in this research and the project relies entirely on publicly available texts, informed consent for data sharing or long-term preservation is not required. The project adheres to the principles of using open-access texts and does not interact with private individuals who would need to provide informed consent.

Other issues

Do you, or will you, make use of other national/funder/sectorial/departmental procedures for data management? If yes, which ones (please list and briefly describe them)?



Yes, the project will follow the data management procedures outlined by Horizon Europe, University of Helsinki, and KU Leuven. These institutions provide detailed guidelines for data security, storage, and sharing that align with the FAIR principles. The project will also follow any specific procedures set by the MECANO project in terms of managing and sharing research data. These procedures ensure that data is stored, shared, and reused according to institutional policies and funding requirements.

HISTORY OF CHANGES		
VERSIO N	PUBLICATION DATE	CHANGE
1.0	05.05.2021	Initial version (new MFF).
1.1	01.04.2022	Reformatted to align with other deliverables templates.

9. Contextual scientometrics—Uncovering and understanding referencing patterns to the ancient canon in modern scholarly discourses

DATA MANAGEMENT PLAN

(To be filled in and uploaded as deliverable in the Portal Grant Management System, at the due date foreseen in the system (and regularly updated).)

⚠ The template is recommended but not mandatory. If you do not use it, please make however sure that you comply with the research data management requirements under Article 17 of the Grant Agreement.)

PROJECT	
Project number:	[project number]
Project acronym:	[acronym]
Project name:	Contextual scientometrics—Uncovering and understanding referencing patterns to the ancient canon in modern scholarly discourses

DATA MANAGEMENT PLAN	
Date:	17/02/2025
Version:	Version 1

Data Summary

Will you re-use any existing data and what will you re-use it for? State the reasons if re-use of any existing data has been considered but discarded.

My main source of data will be a reused dataset, built for the paper Luhmann, J. & Burghardt, M. (2022). Digital humanities—A discipline in its own right? An analysis of the role and position of digital humanities in the academic landscape. *Journal of the Association for Information Science and Technology*, 73(2), 148-171. <https://doi.org/10.1002/asi.24533>

What types and formats of data will the project generate or re-use?

The project will primarily re-use textual data from the afore mentioned corpus. This dataset consists of full-text academic articles in machine-readable format (plain text, .txt format). If the corpus is expanded, new data will follow the same structured format to maintain consistency. The original dataset also includes metadata associated with the articles, such as publication date, discipline classification, country of publication, etcetera.

Other kinds of data such as citation networks, lists of authority control names and derived information such as graphics, will be generated. The non-academic secondment in the Deutsche Nationalbibliothek will provide me as well with a (meta)dataset to employ as a case study.

What is the purpose of the data generation or re-use and its relation to the objectives of the project?

The purpose of re-using this dataset is to conduct further analysis on the evolution of academic discourse across disciplines, particularly in the digital humanities field. By extending the timeframe, the project aims to assess longitudinal trends, thematic shifts, and disciplinary influences over time. The re-use of this extensive dataset aligns with the project's goal of extracting deeper insights without duplicating the effort required to compile such a corpus.

What is the expected size of the data that you intend to generate or re-use?

The corpus consists of 56116 articles with over 299 million tokens. It covers a timeframe of three decades (1990-2019) and features articles from 16 disciplines (statistics, theoretical computer science, information science, mathematics, applied computer science, computational linguistics, political science, sociology, linguistics, digital humanities, art history, musicology, philosophy, history, literary theory and classical studies).

If the corpus is expanded to include more recent publications (beyond 2019), the dataset's size will increase accordingly. The expected growth depends on the availability of articles from relevant sources but could result in a dataset exceeding 60,000 articles and potentially surpassing 350 million tokens.

What is the origin/provenance of the data, either generated or re-used?

The data of the original corpus was "gathered from JSTOR, who—upon our request—provided with the articles as plain text files that were extracted from PDF sources via optical character recognition (OCR). Articles of non-humanities disciplines were mostly retrieved via the CrossRef text mining service as PDF files." (Luhmann & Burghardt, 2022). Any newly generated data will come from similar sources, ensuring continuity in dataset structure and coverage.

To whom might your data be useful ('data utility'), outside your project?

The dataset (and any other interesting output from the project) may be valuable to researchers in digital humanities, bibliometrics, scientometrics, and interdisciplinary studies who are interested in analysing academic discourse trends. Additionally, computational linguists and NLP researchers might find the corpus useful for training language models on academic writing. Scholars studying the evolution of specific disciplines or the impact of digital scholarship on traditional fields may also benefit from this dataset.

FAIR data

Making data findable, including provisions for metadata

Will data be identified by a persistent identifier?

Whenever possible, the generated datasets will be assigned persistent identifiers such as DOIs (e.g., via Zenodo) or GitHub repository links with version control. The re-used dataset from Luhmann & Burghardt (2022) does not have a DOI, so we can publish and generate the identifier for the dataset on its final version once it is extended. The metadata dataset from the Deutsche Nationalbibliothek will be treated similarly, ensuring proper attribution and traceability. Currently, the only output generated by the project (a dataset to build a dictionary of ancient authors, and a list of aliases) lives on GitHub in preliminary versions that should be improved to meet repository standards.

Will rich metadata be provided to allow discovery? What metadata will be created? What disciplinary or general standards will be followed? In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how.

Comprehensive metadata will be provided for all generated and re-used datasets. Metadata will include bibliographic information, data sources, formats, coverage (time period, disciplines), and processing methodologies.

Will search keywords be provided in the metadata to optimize the possibility for discovery and then potential re-use?

Keywords relevant to digital humanities, bibliometrics, citation networks, scholarly communication, and NLP will be included in the metadata. Standardised keywords from controlled vocabularies such as Library of Congress Subject Headings (LCSH) and Getty's Art & Architecture Thesaurus (AAT) will be used where applicable.

Will metadata be offered in such a way that it can be harvested and indexed?

Metadata will be structured to enable machine readability and indexing by search engines, repositories, and research data aggregators. If hosted on GitHub, structured metadata files will be included for web crawlers. If deposited in institutional repositories or Zenodo, metadata will be exposed via OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting) to facilitate discoverability. The Deutsche Nationalbibliothek metadata case study will adhere to their established metadata-sharing practices to ensure compatibility with existing library systems.

Making data accessible

Repository:

Will the data be deposited in a trusted repository?

Data will be deposited in trusted repositories such as Zenodo, GitHub, or institutional repositories (the repository of the University of Leipzig and/or the repository of KU Leuven, as well as repositories of the partner institutions such as the Deutsche Nationalbibliothek) that ensure long-term preservation and accessibility.

Have you explored appropriate arrangements with the identified repository where your data will be deposited?

Datasets will be prepared for deposition in appropriate repositories, ensuring compliance with FAIR principles. If necessary, discussions with repository managers will be held to confirm storage, licensing, and discoverability options.

Does the repository ensure that the data is assigned an identifier? Will the repository resolve the identifier to a digital object?

Zenodo and institutional repositories automatically assign DOIs (Digital Object Identifiers) to uploaded datasets, ensuring persistent access. GitHub repositories can be linked to Zenodo, generating DOIs for software and code versions.

Data:

Will all data be made openly available? If certain datasets cannot be shared (or need to be shared under restricted access conditions), explain why, clearly separating legal and contractual reasons from intentional restrictions. Note that in multi-beneficiary projects it is also possible for specific beneficiaries to keep their data closed if opening their data goes against their legitimate interests or other constraints as per the Grant Agreement.

In general, our data will not require restricted access, as it will not contain sensitive data or copyrighted content. The Deutsche Nationalbibliothek metadata will follow their access policies, which may include licensing restrictions.

If an embargo is applied to give time to publish or seek protection of the intellectual property (e.g. patents), specify why and how long this will apply, bearing in mind that research data should be made available as soon as possible.

Will the data be accessible through a free and standardised access protocol?

Open data will be available through standard access protocols such as HTTP/HTTPS (via Zenodo, GitHub) and OAI-PMH for metadata harvesting.

If there are restrictions on use, how will access be provided to the data, both during and after the end of the project?

In case there were restricted data, access would be granted on request, possibly through institutional agreements, licensing conditions, or research collaborations.

How will the identity of the person accessing the data be ascertained?

Open data will not require authentication. Restricted datasets may require institutional login, ORCID, or agreement to terms of use before access.

Is there a need for a data access committee (e.g. to evaluate/approve access requests to personal/sensitive data)?

No formal data access committee is planned. However, if necessary, requests for restricted data will be evaluated based on legal and ethical considerations.

Metadata:

Will metadata be made openly available and licenced under a public domain dedication CC0, as per the Grant Agreement? If not, please clarify why. Will metadata contain information to enable the user to access the data?

Yes, metadata will be openly available and licensed under CC0 whenever possible, ensuring unrestricted reuse. Metadata will include information on how to access, interpret, and cite the data.

How long will the data remain available and findable? Will metadata be guaranteed to remain available after data is no longer available?

Data will remain available for at least 10 years, following repository policies. Metadata will persist beyond the availability of the data, ensuring future discovery even if the dataset is removed.

Will documentation or reference about any software be needed to access or read the data be included? Will it be possible to include the relevant software (e.g. in open source code)?

Documentation on data formats, processing workflows, and analysis scripts will be provided. If specific software is needed, open-source tools will be referenced, and code will be hosted on GitHub.

Making data interoperable

What data and metadata vocabularies, standards, formats or methodologies will you follow to make your data interoperable to allow data exchange and re-use within and across disciplines? Will you follow community-endorsed interoperability best practices? Which ones?

The project will follow established metadata and data standards to maximize interoperability. Which specific standard to use (Dublin Core, Metadata Object Description Schema, TEI, CIDOC-CRM) will be decided in the moment of sharing the data, trying to adapt to the most appropriate standard according to the content, format and future desired use of our data. The data will be stored in widely-used formats such as CSV, JSON, XML, and RDF to facilitate interoperability across disciplines.

In case it is unavoidable that you use uncommon or generate project specific ontologies or vocabularies, will you provide mappings to more commonly used ontologies? Will you openly publish the generated ontologies or vocabularies to allow reusing, refining or extending them?

If project-specific vocabularies or ontologies are created (e.g., for specialised disciplinary categories), they will be mapped to widely recognised ontologies to maintain compatibility. However, we don't expect this to happen in the frame of our project.

Will your data include qualified references⁶ to other data (e.g. other data from your project, or datasets from previous research)?

Yes, the project will include explicit links to:

- The Luhmann & Burghardt (2022) corpus, acknowledging its role in data reuse.
- Citation networks and authority control datasets to ensure traceability of referenced works.

⁶ A qualified reference is a cross-reference that explains its intent. For example, *X is regulator of Y* is a much more qualified reference than *X is associated with Y*, or *X see also Y*. The goal therefore is to create as many meaningful links as possible between (meta)data resources to enrich the contextual knowledge about the data. (Source: <https://www.go-fair.org/fair-principles/i3-metadata-include-qualified-references-metadata/>)

- The Deutsche Nationalbibliothek metadata case study, linking generated insights to the original dataset.

When applicable, standardised identifiers (DOIs, ORCIDs, Wikidata IDs, VIAF IDs) will be used to ensure data traceability.

Increase data re-use

How will you provide documentation needed to validate data analysis and facilitate data re-use (e.g. readme files with information on methodology, codebooks, data cleaning, analyses, variable definitions, units of measurement, etc.)?

Comprehensive documentation will be provided, including:

- README files describing dataset structure, methodology, and usage.
- Codebooks explaining variable definitions, transformations, and units of measurement.
- Data processing scripts (where applicable) to ensure reproducibility.
- Data cleaning protocols outlining preprocessing steps, filtering criteria, and normalisation techniques.

Will your data be made freely available in the public domain to permit the widest re-use possible? Will your data be licensed using standard reuse licenses, in line with the obligations set out in the Grant Agreement?

Most datasets will be made openly available under standard reuse licenses, such as CC0 (Public Domain Dedication; for metadata and non-sensitive datasets) and CC-BY 4.0 (Attribution Required; for datasets requiring citation).

Will the data produced in the project be useable by third parties, in particular after the end of the project?

All openly available data will remain accessible for reuse, with metadata ensuring long-term discoverability. Version-controlled repositories will preserve datasets and allow future updates or expansions.

Will the provenance of the data be thoroughly documented using the appropriate standards?

Provenance will be documented following PROV-O (W3C Provenance Ontology) for tracking data derivation, DCMI Metadata Terms for source attribution and standard citation practices ensuring transparency in data reuse.

Describe all relevant data quality assurance processes.

- Validation procedures: automated checks for data completeness and consistency, manual cross-validation of derived information (e.g., author lists, citation networks).
- Error handling: outlier detection and anomaly checks, documentation of known limitations and uncertainties in data collection.
- Reproducibility checks: Availability of analysis scripts to replicate findings, encouragement of community feedback for validation.

Further to the FAIR principles, DMPs should also address research outputs other than data, and should carefully consider aspects related to the allocation of resources, data security and ethical aspects.

Other research outputs

In addition to the management of data, beneficiaries should also consider and plan for the management of other research outputs that may be generated or re-used throughout their projects. Such outputs can be either digital (e.g. software, workflows, protocols, models, etc.) or physical (e.g. new materials, antibodies, reagents, samples, etc.).

Beneficiaries should consider which of the questions pertaining to FAIR data above, can apply to the management of other research outputs, and should strive to provide sufficient detail on how their research outputs will be managed and shared, or made available for re-use, in line with the FAIR principles.

Beyond data management, this project will generate a range of additional research outputs, including digital resources such as software scripts, workflows for data processing, analytical models, and data visualisations. These outputs will be developed following the FAIR principles to ensure findability, accessibility, interoperability, and reusability. Software tools and scripts, primarily used for corpus analysis and citation network visualisation, will be openly shared through GitHub and linked to Zenodo to provide persistent identifiers and clear version control. Workflows and protocols for data processing will be documented in README files, ensuring that third parties can replicate or modify the analytical approaches used in the project. Where project-specific methodologies are employed, they will be mapped to standard approaches whenever possible to ensure broad compatibility.

For non-digital outputs, such as reports and presentations, the project will adhere to open-access policies to maximize dissemination. Where applicable, derived visualisations and analytical models will be shared in an interactive format, ensuring accessibility for both academic and non-academic audiences. The project will also make use of institutional and sectorial guidelines for research data and software management, ensuring alignment with best practices for sustainability and long-term impact.

Allocation of resources

What will the costs be for making data or other research outputs FAIR in your project (e.g. direct and indirect costs related to storage, archiving, re-use, security, etc.) ?

The costs associated with ensuring that data and other research outputs adhere to FAIR principles will include expenses related to secure data storage, long-term archiving, repository fees, metadata curation, and documentation. In general we will choose free repositories, so there are no extra induced costs.

How will these be covered? Note that costs related to research data/output management are eligible as part of the Horizon Europe grant (if compliant with the Grant Agreement conditions)

Who will be responsible for data management in your project?

I will be responsible of the management of the data generated by my own project.

How will long term preservation be ensured? Discuss the necessary resources to accomplish this (costs and potential value, who decides and how, what data will be kept and for how long)?

Long-term preservation will be ensured through a combination of trusted repositories and institutional archiving services, with decisions on retention periods made based on dataset value, usage metrics, and compliance with funder mandates. Metadata and documentation will remain available beyond the lifespan of the project, even if certain datasets are no longer actively maintained.

Data security

What provisions are or will be in place for data security (including data recovery as well as secure storage/archiving and transfer of sensitive data)?

To ensure data security, the project will implement strict protocols for data storage, backup, and recovery. Regular backups will be scheduled to prevent data loss, with redundant copies maintained in geographically separate locations to ensure recovery in case of system failure.

Will the data be safely stored in trusted repositories for long term preservation and curation?

All research data will be stored in trusted repositories that support long-term preservation and secure archiving. Platforms such as Zenodo and institutional data repositories will be used to ensure persistent access while maintaining compliance with data protection regulations. For datasets that require controlled access, authentication mechanisms such as institutional logins or ORCID verification will be employed to regulate data retrieval and track usage. During data transfer, secure file exchange protocols will be used to protect against interception or unauthorized modifications.

Ethics

Are there, or could there be, any ethics or legal issues that can have an impact on data sharing? These can also be discussed in the context of the ethics review. If relevant, include references to ethics deliverables and ethics



chapter in the Description of the Action (DoA).

Ethical considerations will be central to the project's data management strategy, particularly concerning data sharing and legal constraints. The metadata dataset provided by the Deutsche Nationalbibliothek will be managed in accordance with institutional data policies, ensuring compliance with ethical and legal obligations. Ethical approval will be sought where necessary, with documentation included in the relevant ethics deliverables and sections of the Description of the Action (DoA). The project will also adhere to GDPR and national data protection regulations to safeguard privacy and ensure responsible data handling.

Will informed consent for data sharing and long term preservation be included in questionnaires dealing with personal data?

Other issues

Do you, or will you, make use of other national/funder/sectorial/departmental procedures for data management? If yes, which ones (please list and briefly describe them)?

HISTORY OF CHANGES		
VERSIO N	PUBLICATION DATE	CHANGE
1.0	17.02.2025	Initial version (new MFF).

10. A democratic turn? Uncovering and understanding references to Graeco-Roman antiquity in 20th-century French public discourse

This DMP was created in RIS

General information

1.1 Title of this DMP

A democratic turn? Uncovering and understanding references to Graeco-Roman antiquity in 20th-century French public discourse

1.2 Summary of research proposal

This PhD project will explore the use of Graeco-Roman antiquity as a reference culture in the late 19th- and early 20th-century public discourse, in France. It will use reference mining as a tool to develop a longitudinal overview of ancient canon formation as negotiated through journalistic writings. By collecting references to antiquity from a corpus of digitised press publications and categorising them, it will create an instrument to assess existing hypotheses and uncover new patterns related to the reception of classical antiquity. It will also evaluate the robustness of ancient canons in the 'age of the masses.'

1.3 University, faculty, institute, and relevant RDM policy

Radboud University, Faculty of Arts, Radboud Institute for Culture & History (RICH)

Link to the Research Data Management policy of Radboud University:

<https://www.ru.nl/en/about-us/policies-and-regulations/research-data-management/guidelines-for-research->

- 1.4 If applicable: What are the project number, funder, and funder ID? Project 101120349 - MECANO, funded by the European Research Executive Agency (REA), under the powers delegated by the European Commission.

1.5 What is the (expected) start and end date of the project?

Start date: 01/09/2024 - Expected end date: 31/08/2028

- 1.6 Which researcher(s) and/or relevant parties are involved in the research project, and what are their roles regarding data management?

Involved in writing and adjusting the DMP:

Marin-Marie Le Bris (Radboud Universiteit, PhD candidate)

Prof. dr. Maarten De Pourcq (Radboud Universiteit, main supervisor/promotor)

Prof. dr. Manuel Burghardt (Universität Leipzig, supervisor)

Dr. Shari Boodts (Radboud Universiteit, MECANO data officer)

Dr. Roel Smeets (Radboud Universiteit, advisor)

Involved in data collection and analysis:

Marin-Marie Le Bris (Radboud Universiteit, PhD candidate)

Prof. dr. Maarten De Pourcq (Radboud Universiteit, main supervisor/promotor)

Prof. dr. Manuel Burghardt (Universität Leipzig, supervisor)

Dr. Roel Smeets (Radboud Universiteit, advisor)

Luisa Ripoll Alberola (Universität Leipzig, PhD candidate)

Involved in data storage during research:

Marin-Marie Le Bris (Radboud Universiteit, PhD candidate)

Involved in long-term data archiving and sharing after the project, including transfer of data management roles:

Marin-Marie Le Bris (Radboud Universiteit, PhD candidate)

Dr. Shari Boodts (Radboud Universiteit, MECANO data officer)

1.7 Who is the rights holder of the data that you will collect and/or generate during the project?

Radboud University is the rights holder of my research data.

1.8 Do you need to create or sign any agreements on data management during the project

(such as consortium agreements, data use agreements, and non-disclosure agreements)? If yes, please specify the agreements.

No, no such agreements are required for my research.

1.9 Did you consult a local expert (e.g. your institute's data steward or the RDM support team) when writing this DMP? If yes, specify who you consulted and specify the date of consultation.

On the 28th of August 2024, during Radboud Universiteit's (RU) Graduate School of Humanities' (GSH) Introduction Days, I participated in the workshop 'Research Data Management' (RDM), conducted by Henk van den Heuvel (Head of the Humanities Lab and Research Data Manager at the Faculty of Arts) and Sanna van Roosmalen (Policy Officer Research/Valorisation).

A specific session dedicated to the writing and updating of a Data Management Plan (DMP) was provided by KU Leuven's RDM team to all PhD candidates from the MECANO doctoral network on the 20th of September 2024, at the Kortenbergh Abbey in Belgium.

In addition, on the 30th of September 2024, I received feedback regarding my initial RDM draft from RU's data stewards of the Faculty of Arts and the Digital Competence Centre. I have subsequently done my best to implement their suggestions.

Data collection

2.1 Will existing data be (re-)used for this research project? If yes, please specify the data, their source and the terms of re-use.

Yes, existing data will be (re-)used during the project. The data, their source and terms of re-use are as follows: Aside from primary and secondary literature (which will be cited and listed in my dissertation's bibliography), the existing data that will be (re-)used during the project will consist of:

- (1) 19th- and 20th-century newspaper issues, in the form of text files (TXT) obtained from Gallica, the Bibliothèque nationale de France's (BnF) digital library. The "non-commercial use of [these] documents is free of charge, in accordance with the legislation in force, and in particular with the requirement to mention the source as BnF or

Bibliothèque nationale de France," see Gallica's 'Research assistance' page: <https://gallica.bnf.fr/accueil/en/html/research-assistance>

- (2) Python scripts (PY) to retrieve the newspaper issues mentioned above, made available by Julien Schuh and Pierre-Carl Langlais in the form of a programme called 'Pyllica', under a GNU General Public License v.2.0 (see Pyllica's page on GitHub: <https://github.com/Dorialexander/Pyllica/blob/master/LICENSE>).
- (3) Readership-related data to assess the nature and the size of the audience of selected press publications, digitally reproduced in CSV/XML/JSON/TXT formats (when necessary), obtained from close reading of sources such as: Pierre Albert et al., *Histoire générale de la presse française*. Tome III: De 1871 à 1940, ed. Claude Bellanger et al., vol. 3 (Paris: Presses universitaires de France, 1976). These sources will be clearly mentioned in the corresponding files, and properly cited every time the data are used qualitatively or quantitatively.
- (4) Metadata associated with the periodicals that will be part of the newspaper corpus mentioned in (1), (2) and (3), retrieved via Gallica's 'API Document' (<https://api.bnf.fr/fr/api-document-de-gallica>) in XML format.
- (5) Lists, dictionaries and lexica of ancient authors, works, citations, concepts, locutions etc. in French, Latin and Greek. These will be compiled from primary and secondary literature *on* and *from* Graeco-Roman antiquity and, when needed, digitally reproduced as CSV/XML/JSON/TXT files. Sources will then be clearly mentioned, and properly cited every time the data are used qualitatively or quantitatively.

2.2 Will new data be generated within the research project? If yes, please specify the collection process and the data that will be generated, including file formats.

Yes, new data will be generated during the project. The collection process and the data that will be generated are as follows:

To identify direct or indirect references to Graeco-Roman antiquity in French newspaper issues of the late 19th- and early 20th-century, I will programme a 'reference extractor', i.e. a tool to match the author names, works, citations, concepts, locutions etc. mentioned in (5) of question 2.1, to the journalistic writings mentioned in (1) of question 2.1. I will then conduct a series of statistical analyses on the collection of matched items to unveil meaningful trends and relationships.

The data generated during the project will therefore consist of:

- (a) The Python scripts (PY) used to programme the 'reference extractor' and run the statistical analyses.

- (b) The lists of references to Graeco-Roman antiquity found in the newspaper corpus (in CSV/XML/JSON/TXT formats).
- (c) The collection of excerpts and journalistic writings containing references from (b) (in TXT format).
- (d) A number of statistical indicators, tables and other results, drawing on readership-related data and relevant metadata to uncover and describe referencing patterns found in the newspaper corpus. Those results will be saved in CSV/XML/JSON/TXT files, depending on the nature of the methods implemented.

Note that while relying on pre-existing data, the extensive newspaper collection mentioned in (1) of question 2.1. (i.e. which will comprise *not only* publications referencing antiquity but *all* examined journalistic writings) may also be presented as 'new data', given that it will most likely be composed according to project-specific criteria, and will therefore differ from any currently available corpus.

The same holds true for the compilation of lists, dictionaries and lexica that will be used as a proxy for GraecoRoman antiquity to look for references in press publications (see (5) of question 2.1.).

2.3 What is the total expected size of your data?

Given that the data used and generated in this project will mostly consist of *text*, their total size should not exceed the 1GB - 10 GB range. It may even be smaller than 1GB. However, I will likely generate some tables to visually display the output of the statistical analyses carried out on the references found in the journalistic corpus. These images will be saved as either JPEG or TIFF files.

Personal data

- 3.1 Do any of the project's data allow identification of a person? In other words, are you working with personal data? List all types of data in your dataset which could be used for identification.

Do not forget about

- data that you use for participant recruitment, contact information, the key file of pseudonymised data, etc.
- data that can lead to identification when combined (e.g. place of residence and job description in some cases) - personal data that you do not specifically ask for, but participants may provide in response to an open question in a questionnaire or during an interview
- Other, namely:

The articles from the newspaper issues I will be using have all been written by journalists working for periodicals.

While some may mention the names of their authors, such information is already publicly available. I expect most articles to be anonymised and presented as the result of a collective endeavour (the newspaper). The same goes for the opinions, beliefs and ideas - sometimes political, philosophical or religious - conveyed in those articles.

- 3.2 Do your data contain special categories of personal data? List all categories and specify the data.

Be aware that special categories of personal data are subject to strict legal conditions.

- Political opinions, namely:

See the remark above (question 3.1.)

- Religious or philosophical beliefs, namely:

See the remark above (question 3.1.)

3.3 List all third party software or online services which will have access to the (special categories of) personal data during the collection and/or analysis phase.

- Not applicable, I do not work with personal data.

3.4 Will you anonymise or pseudonymise the data in order to protect the privacy of your participants? Please specify how or why not.

Not applicable, I do not process personal data.

3.5 Do you need approval from an ethics committee for your project? Please explain why (not). No, I do not need approval from an ethics committee for my research, because:

I will not be gathering new personal data and only build on existing, publicly available and openly accessible data.

3.6 Do you need to use an informed consent procedure?

No, I do not work with human participant data and therefore an informed consent procedure is not necessary.

Storing and sharing during research

4.1 Explain where you will store your data during research.

Name *all* storage facilities and/or devices you will use during research, even when only used temporarily.

- M365 Teams

I will follow my institute's policy and store my data in a Microsoft 365 Team with a RU account during the research phase of my work. Safe and secure storage is guaranteed by the IT security and safety protocols. The data are automatically backed up on a regular basis.

This will allow me to share the data with both internal and external colleagues involved in the project.

4.2 How will you share your data during research and with whom? Specify whether they are Radboud researchers or from a different Dutch or international institute.

Be aware that sharing personal data should be kept to a minimum. Furthermore, not all tools are suitable for sharing personal data, see the 'i'-button for details on the options below.

- I will follow my institute's policy and use M365 Teams to share my data with:

Internal colleagues (from Radboud Universiteit), namely:

Prof. dr. Maarten De Pourcq (Radboud Universiteit, main supervisor/promotor)

Dr. Shari Boodts (Radboud Universiteit, MECANO data officer) Dr. Roel Smeets (Radboud Universiteit, advisor)

External colleagues (from other international institutions), namely:

Prof. dr. Manuel Burghardt (Universität Leipzig, supervisor)

Luisa Ripoll Alberola (Universität Leipzig, PhD candidate)

Long-term archiving and reuse

5.1 Where will you archive your data (including raw data, metadata, and documentation) for at least 10 years for the sake of scientific integrity?

This may be done in an internal or public archive. Be aware that personal data that are not necessary for answering the research question (e.g. contact details of participants) must be deleted as soon as possible and should not be archived, regardless of the choice of archive.

- Radboud Data Repository

I will follow my institute's policy and archive the research data associated with my publication (including raw data, metadata, and documentation) in the Radboud Data Repository for a minimum of 10 years.

5.2 Will you make (parts of) your research data publicly available for re-use and replication purposes? Please specify where and when and whether any restrictions or embargoes apply. If you are not making your data publicly available, provide a valid reason for not sharing your data.

Be aware that you are not allowed to share personal data publicly unless you have permission from your participants through informed consent.

- Radboud Data Repository

Yes, I will follow my institute's policy and make the parts of my research data that are suitable for publication publicly available via the Radboud Data Repository. This includes a persistent identifier (DOI), metadata (Dublin Core/Datacite), documentation, and one of the available licences.

In the Radboud Data Repository, I will make the following data publicly available:

- The comprehensive corpus of 19th- and 20th-century newspaper issues serving as a proxy for French public discourse, retrieved (in TXT format) from Gallica (see (1) of question 2.1);
- The compilation of lists, dictionaries and lexica used as a proxy for Graeco-Roman antiquity when looking for references in the abovementioned periodicals (in CSV/XML/JSON/TXT formats) (see (5) of question 2.1);
- The excerpts and journalistic writings found to be containing references to antiquity (in TXT format) (see (c) of question 2.2);
- The final collection of references found in the newspaper corpus (in CSV/XML/JSON/TXT formats) (see (b) of question 2.2);
- The Python scripts (PY) used to programme the 'reference extractor' and to conduct statistical analyses on identified references to antiquity and their contexts (see (a) of question 2.2);
- The results of the statistical analyses conducted on items matched by the reference extractor (in CSV/XML/JSON/TXT formats) (see (d) of question 2.2).

All of these will be stored under the following license: CC0.

The data will be made available as soon as the articles/thesis are published.

5.3 How will you ensure that your research data will be archived and/or published in a FAIR manner?

- Radboud Data Repository

I will use the Radboud Data Repository to archive and/or publish my data. My data will comply with the FAIR principles in the following way:

My data will be Findable via the Radboud Data Repository, which is indexed by search engines on the internet. My data includes rich metadata and has a persistent identifier (DOI).

My data will be Accessible as well, since the Radboud Data Repository uses an open internet protocol, including clear authorisation procedures.

My data will be Interoperable by the use of standards for metadata (Dublin Core/Datacite), documentation, preferred formats and, if existing, standard (domain-specific) vocabularies.

My data will be Reusable via the Radboud Data Repository, including rich metadata, documentation for reuse, and a clear licence.

Costs

6.1 Do you foresee extra costs for data management that are not covered by your institute?

All the costs are covered by my institute.

My institute provides me with computers, software, and storage space during research and for long-term storage.