DMP title

Project Name My plan (Internal Funds DMP) - DMP title
Project Identifier C3/21/005
Grant Title C3/21/005
Principal Investigator / Researcher Mark Depauw
Institution KU Leuven

1. General Information

Name of the project lead (PI)

Mark Depauw

Internal Funds Project number & title

C3/21/005 Linking traditional scholarship of the Ancient World with digital tools through Artificial Intelligence

2. Data description

2.1. Will you generate/collect new data and/or make use of existing data?

- Generate new data
- Reuse existing data

2.2. What data will you collect, generate or reuse? Describe the origin, type and format of the data (per dataset) and its (estimated) volume. This may be easiest in a numbered list or table and per objective of the project.

Type of data	Format	Volume	How created
PDF's articles, books, ...	.pdf	++	Scraped from websites and through contact with publishers
Metadata of PDF's	MySQL database	++	Extracted from PDF metadata and added at time of scraping
Extracted strings	JSON	++	Relevant sections in the text of the pdf's and their positions, by setting manual ground truth, and rule-based and machine learning system. a learning set of data generated from PDFs (Articles, books, etc.) where the valuable information is extracted along with its context. A span of text is generated for each valuable information in order to feed the AI model in the training process.
Evaluations and Interpretations of the extracted strings	MySQL	++	In first instance rule-based, later also machine-learning algorithms, manually corrected afterwards
Set of connections between TM database and extracted relevant strings	MySQL	++	Result of all the previous manipulations

3. Ethical and legal issues

3.1. Will you use personal data? If so, shortly describe the kind of personal data you will use. Add the reference to the file in KU Leuven's Record of Processing Activities. Be aware that registering the fact that you process personal data is a legal obligation.

We will not process personal data directly. Indirectly we will occasionbally work with the names of the individuals who have published certain editions, but this is only in public capacity.

3.2. Are there any ethical issues concerning the creation and/or use of the data (e.g. experiments on humans or animals, dual use)? If so, add the reference to the formal approval by the relevant ethical review committee(s).

No ethical issues.

3.3. Does your research possibly result in research data with potential for tech

transfer and valorisation? Will IP restrictions be claimed for the data you created? If so, for what data and which restrictions will be asserted?

The resulting data have valorisation potential, as they will allow to set up a bibliographic database with mentions of each source, but also connecting the source in the pdf through a hyperlink with existing academic tools. The annotations are therefore our IP, and will only be accessible to us and the authorised users of the tool.

3.4. Do existing 3rd party agreements restrict dissemination or exploitation of the data you (re)use? If so, to what data do they relate and what restrictions regarding reuse and sharing are in place?

The data in the PDF's are public, and thus there is in our opinion no problem if we annotate them. We will discuss with publishers about making new material available for annotation, possibly at an early stage.

4. Documentation and metadata

4.1. What documentation will be provided to enable understanding and reuse of the data collected/generated in this project?

The metadata of the pdf's will be stored in a MySQL database 'pdf_annotations'. The following tables have been implemented: publisher, journal, publication, scraped_pubs, scraped_documents and pdf_annotations. The MySQL database is maintained on the Trismegistos server, which has a strict backup schedule (once every hour), with regular backups on other machines as a second backup layer. We will write a document explaining the structure of the information which will be added to GitHub.

4.2. Will a metadata standard be used? If so, describe in detail which standard will be used. If not, state in detail which metadata will be created to make the data easy/easier to find and reuse.

We will use DOI's for the scraped publications (scraped_documents). The publications will also be linked to Trismegistos publication table.

The metadata used for each article (PDF, Book, etc.) is structured such that the DOI would be the primary key followed by information like the author, the title of the article and the type of the article.

5. Data storage and backup during the project

5.1. Where will the data be stored?

The data will be stored on the Trismegistos server, which is managed through a regular backup service with a second layer on external machines. The code is made available to all members of the project on GitHub.

5.2. How will the data be backed up?

See previous question.

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

We currently have enough storage and backup capacity for the project through existing procedures.

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

The data will become part of the Trismegistos research environment, the costs of which are covered by the Trismegistos+ KU Leuven Core Facility, which is also a member of the CLARIAH ERIC environment.

5.5. Data security: how will you ensure that the data are securely stored and not accessed or modified by unauthorized persons?

The server used for the data of this project is protected by strong password connections and can only be accessed from within the KU leuven network.

6. Data preservation after the end of the project

6.1. Which data will be retained for the expected 10 year period after the end of the

project? If only a selection of the data can/will be preserved, clearly state why this is the case (legal or contractual restrictions, physical preservation issues, ...).

All resulting annotations will be kept as part of a new table in the Trismegistos enviroment and will be fully integrated in the rest of the project. For the preservation of the pdf's we will explore a model depending on coopertaion with the publishers.

6.2. Where will these data be archived (= stored for the long term)?

The data will be stored on the university's central servers (with automatic back-up procedures) for at least 10 years, conform the KU Leuven RDM policy.

6.3. What are the expected costs for data preservation during these 10 years? How will the costs be covered?

The costs will be covered by the Trismegistos+ KU Leuven Core Facility in the framework of Trismegistos.

7. Data sharing and re-use

7.1. Are there any factors restricting or preventing the sharing of (some of) the data (e.g. as defined in an agreement with a 3rd party, legal restrictions or because of IP potential)?

The data will become part of a module with subscribers-only restricted access within the Trismegistos website. We will develop a business model and resultingly negotiate with publishers about the sharing and spread of data.

7.2. Which data will be made available after the end of the project?

The extracted relevant sections of the pdf's and their interpretation as references to sources included in the Trismegistos Text or Trismegistos Authors databases will be made available either as part of the Trismegistos subscriptions or in a separate module.

7.3. Where/how will the data be made available for reuse?

On the basis of a subscription (see above). We will explore whether to release the source code on GitHub.

7.4. When will the data be made available?

- Immediately after the end of the project

We will provide subscribers' access to the annotations as soon as these are available.

7.5. Who will be able to access the data and under what conditions?

Access will only be possible for full subscribers to Trismegistos, with possible an extra subscription.

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

As the data will be shared on the basis of a subscription, it is this subscription wchich will cover the costs.

8. Responsibilities

8.1. Who will be responsible for the data documentation & metadata?

The data documentation will be the joint responsibility of the project members: Frederic Pietowski for the management of the scraped documents and their metadata, Lamine Benrais and Bart Thijs for the code and annotation environment, and Mark Depauw for the incorporation into Trismegistos and the general coordination.

8.2. Who will be responsible for data storage & back up during the project?

Data storage and backup will be maintained within the Trismegistos project, with Frederic Pietowksi as coordinator.

8.3. Who will be responsible for ensuring data preservation and sharing?

The long-term integration and functionality of the resulting data is the responsibility of Mark Depauw.

8.4. Who bears the end responsibility for updating & implementing this DMP?

The end responsibility for updating and implementing the DMP is with the supervisor (promotor).