DMP du projet "Hospitam"

Plan de gestion de données créé à l'aide de DMP OPIDoR, basé sur le modèle "ERC DMP" fourni par Conseil européen de la recherche (European Research Council, ERC).

Plan Details

Plan title	DMP du projet "Hospitam"
Language	fra
Creation date	2019-11-18
Last modification date	2020-04-24
Identifier	4223

Project Details

Project title	Hospitam
Abstract	
Funding	• Conseil européen de la recherche (European Research Council, ERC) :

Research outputs :

1. Default research output (Jeu de données)

Contributors

Name	Affiliation	Roles
Lauriane Locatelli		• Coordinateur du projet • Personne contact pour les données • Responsable du plan

DMP du projet "Hospitam"

Summary

Project Acronym

HospitAm

Project Number

/

Provide a dataset summary *(Several datasets may be included into a single DMP)*

/

FAIR data and resources

1. Making data findable

The data can be found via the **Handle Nakala**, and via **Isidore** which retrieve informations from **Nakala**. Metadata will follow the Dublin Core standard. The volume of each data will only be up to a **few GB**. The creation of an easily accessible presentation page **or landing page describing the dataset** is currently contemplated. The landing page will be on the Project's website.

The most relevant literary sources relating to the analysis of hospitality in Antiquity and the crises of hospitality in Antiquity have been examined. A corpus of 6 ancient authors has been determined: Dionysos of Halicarnassus, Plutarch, Apollonius of Tyana, Basil of Caesarea, John Chrysostomus, Libanios. A body of hospitality passages from the Bible is also currently processed. The Bible is read to isolate the hospitality verses. The lexical field of hospitality and the crisis of hospitality has been researched.

Afterwards, the texts in digital versions will be recovered from a site like Treebank Data Perseus (https://github.com/PerseusDL/treebank_data) or else First One-Thousand Years of Greek Project (https://opengreekandlatin.github.io / First1KGreek /)., Then the texts will be lemmatized using a **python script produced by Marianne Reboul for lemmatization purposes**. Lemmatized texts would ideally be used. Unfortunately, few lemmatized texts exist – retrieve non-lemmatized texts and then lemmatize them turned out to be easier to achieve. Regarding the script code, it will be documented using a **Jupiter Notebook** and deposited on **Gitlab**.

We are working in collaboration with the **GREgORI project of UCLouvain** in the person of Bastien Kindt. The GREgORI dissemination tools are **Open Source** and provide expertise in the lexical labeling of corpora (lemmas, morphosyntactic categories, grammatical analyzes), in classical and oriental languages (preparation of corpora, processing, exploitation, dissemination, etc.); dissemination tools - static matches ("word" matches in PDF or other format) or dynamic matches ("word" matches or "word sequences" on WEB interfaces) - available in Open Source. A first test was carried out on the lexical fields of hospitality on PROCOPE. The GREgORI project so far benefits from the corpus of PROCOPE and ZOSIME - already treated, they could already serve as a basis for tests and concrete evaluations. The results of these tests will be accessible via an online interface.

New data will be produced by collecting non-lemmatized texts from the corpus and then lemmatizing them using a Python script already in possession of one of the project's collaborators (Marianne Reboul, ENS Lyon). It has been successfully tested on a sample corpus. This mission will be accomplished by a post-doctoral fellow in digital humanities. Then, a form-

matching will be carried out with the help of a project partner (Bastien Kindt, UCLouvain). The corpus of ancient authors in connection with a biblical passage relating to hospitality will be created through the online tool **Biblindex**. BiblIndex is an index of biblical citations and allusions found in Eastern and Western patristic literature. Intended to facilitate and renew the study of the reception of scriptural texts, BiblIndex can also be used as a synopsis of online Bibles (Biblical tools) or as an index of patristic works (Patristic tools). The export will be done in html. The preparatory work has already been done. Online tools from hypertext environments (Hodoi elektronikai, Itinera electronica) will be used. It is an enhancement of existing data. The feasibility has therefore been checked using sample corpora. The data produced will be in **.txt** format.

The resulting data will be as follows:

- Textual documents of ancient texts neither lemmatized nor lemmatized in Greek language (ancient Greek alphabet with accent, **Unicode** format) in **.txt** grouped in files by authors, that is to say 6 files of maximum 30 MB, 30MB being the required volume for files on Libanios. At present, we have 119 .txt files (16 + 103) for the lemmatizations of Diosysos of Halicarnassus (16 doc .txt) and Libanios (103 doc .txt).
- Textual documents of biblical passages linked to hospitality in Greek, Latin and English in .odt format (a free format), ie 73 documents as there are 73 books in the Bible, with individual files of around 500Kb
- **xml documents** for the results Biblindex of each passage studied with its references from ancient authors, ie between 3000 and 4000 files as the Bible has 73 books - for each book, we believe that there may be 50 passages (for the moment approximately 30 for Genesis and for Matthew), 73x50 = 3650 files.
- **Zotero bibliography**, to manage bibliographic references on the subject, maintained by the project leader.
- **Spreadsheet in .csv** containing the list of authors.

The data will be accompanied by readme.txt documentation detailing the methodology used to collect data from Biblindex. It is possible to create a **didactic web page** referring to the various resources to facilitate reusing the data for the users who desire to do so.

NAMING: Regarding naming conventions, the names of biblical books are given in English. The number of a biblical book will be separated from the verse number by an underscore. As for passages, the beginning and the end of the passage will be separated by a dash. For example, Matthew26_7-13 for Matthew book 26, passage from verse 7 to verse 1. File names will not include special characters, no spaces.

Each file name ends with an indication of the date and time of the backup in American format, for example **Mattew26_7-13_01-13-2020_12-24.**

Regarding the files, the folder of the six Greek authors will be separated from the work file in connection with the Bible. The folder “6_auteurs_grecs” will be made up of 6 sub-folders, corresponding to each of the 6 authors, namely: Dionysos_of_Halicarnassus, Plutarch, Apollonius_of_Tyana, Basil_of_Caesarea, Chrysostomus and Libanios.

Structure of the database: the database will include at least two tables: an author table and a text table. The database will be supplemented by hand-made CSVs.

---

2. Making data openly accessible

The data will be accessible via **ISIDORE** and will be stored on a **Huma-Num Box**. The quality of the data will be checked through validations decided during a **fortnightly validation meeting** (twice a month). These meetings will bring together the project leader, the post-doctoral fellow in digital humanities and the design engineer. Tools allowing to check the links of the database which refer to hypertext environments, such as for example **LinkChecker** (a free software) will be used. This tool allows you to check for broken links in HTML documents. LinkChecker is a command line python tool which allows you to browse a site by following the links. It provides a summary (number of warning, number of errors) and is configurable to suit the project’s needs. It will be used by the project design engineer. In the database spreadsheet **a column will indicate the status of a data**: "in progress / validated on DATE". The data will be subject to intellectual validation, expert validation by the project leader.

Regarding **versioning management**, a version management software will be used: **Git**. Git is a free software.

Regarding data access, the data will be accessible respecting the principles of Open Science. The data used in this project are not sensitive data. Our data are ancient texts free of rights like the bible or the text of ancient authors, not belonging to publishing houses. We do not have personal data

Literary and biblical references will be freely accessible through the database. The files containing the lemmatized texts will also be freely accessible in order to promote Open Access. The data obtained using the Biblindex tool will meet the same principles as the Biblindex site, and its data will be freely accessible as long as the Biblindex site is freely available as well. Reuse will be granted in accordance with the **CC BY NC license (Creative Commons)** wich is free and guarantees the protection of copyright and the **Etalab license**, which is a free French license.

We reflect on the question of copyright concerning the structure of the database, aware that this question deserves to be raised.

---

3. Making data interoperable

The data will be more easily interoperable as we will use the **Dublin Core standards**. Metadata can be harvested using the OAI-PMH protocol. The **OAI-PMH** protocol for the interoperability of open archives is based on a simple **Dublin Core** minimal record in XML.

The data is disseminated via Nakala and preserved via the **Huma-Num Box.** Huma-Num Box will be used for archiving.The data is described in the DublinCore standard to guarantee interoperability. Regarding storage, the metadata is saved in Nakala and the data that is described by this metadata stored in the HumaNum Box. Isidore harvests Nakala and makes this metadata accessible. The use of open and standardized formats will be encouraged. The following tools will be used to control the **quality of file formats**: **the FACILE tool of CINES** (National Computer Center for Higher Education) and the W3C website for the validation of xml formats. The data will be subjected to technical validation by the design engineer. A **controlled vocabulary** (that of the Thesaurus Linguae Graecae) will be used

---

4. Increase data reuse

The data will be well described to be reusable. We use **controlled vocabulary** to standardize the vocabulary about hospitality, but also to standardize the names of texts by Greek authors. We will use the **community standard of the Linguae Graecae Thesaurus**.

The translation of Libanios will temporally be unaccessible as it is the result of an ongoing seminar work which may be published. Upon publication, the translation of Libanios will be made available.

There will be a **website developed by the MOM** (Maison de l'Orient et de la Mediterranean) and **hosted by the HiSoMA** laboratory. Part of the information for visibility and restricted access to specific groups: on the one hand the members of the project and on the other hand the other members of the laboratory

Our project is not concerned with ethical issues

---

5. Allocation of resources and data security

The project will benefit from a range of tools and services like **Nakala** for data exposure. Storage will be delegated to **Huma-Num** and preservation will be managed by **Huma-Num**. The aim is to successfully preserve the document and the information it contains for the duration of the project.

The data will be retrieved and shared by the repository in a trusted data warehouse, such as **Nakala.**

Data such as lemmatized texts can be reused in the context of other research projects or for teaching.

Data will be accessible via **ISIDORE. ISIDORE** consults the data stored in Nakala. A standard metadata structure like the Dublin Core, which will make the metadata interoperable (rending it possible to connect it to other existing warehouses) and make it harvestable by specialized services such as Isidore.

The data will have **a unique identifier**: a **Nakala Handle.**

The data manager will be the recruited engineer. He will manage: the repository on **Nakala**, the Zotéro bibliography and the evolution of the database structure. Data entry and metadata production will be carried out by the project researchers. The engineer will take care of data quality, storage and backup, as well as data archiving and sharing. The engineer will be responsible for implementing the data management plan. Data entry and data production will be carried out by the post-doctoral fellow. He will also be responsible for data harmonization.

The data management plan will be revised and updated regularly (with at least 3 versions during the project) at a meeting bringing together the PI, the post-doctoral student and the engineer. The contact person at the end of the ERC concerning data management will be the project leader.

Data management will occupy about 25% of the engineer's time, or about 8:45 a week. The resources necessary for the dissemination of data are: data warehouses like Nakala.

The costs to be expected apart from the human resources costs are, for example, the purchase of external hard drives (2 2TB drives, between 500 € and 650 € for a 2TB SanDisk hard drive.