# Hospitam

*Plan de gestion de données créé à l'aide de DMP OPIDoR*

**Créateur du PGD :** Lauriane Locatelli

**Affiliation du créateur principal :** ENS de Lyon

**Modèle du PGD :** ANR - Modèle de PGD (français)

**Dernière modification du PGD :** 24/04/2020

**Financeur :** IDEL LYON (ELAN ERC)

**Résumé du projet :**

Hospitalités dans l’Antiquité méditerranéenne

**Chercheur Principal :** Lauriane Locatelli

**Contact pour les Données :** Lauriane Locatelli

# Hospitam

## 1. Description des données et collecte ou réutilisation de données existantes

De nouvelles données seront produites en collectant les textes non lemmatisés du corpus puis en les lemmatisant à l'aide d'un script en Python qu'un des collaborateurs du projet possède déjà (Marianne Reboul, ENS Lyon). Il a été testé avec succès sur un corpus échantillon. Cette mission sera accomplie par un post-doctorant en humanités numériques. Ensuite, une concordance-formes sera réalisée grâce à l'aide d'un partenaire du projet (Bastien Kindt, UCLouvain). Le corpus des auteurs antiques en lien avec un passage biblique relative à l'hospitalité sera constitué à partir de l'outil en ligne Biblindex. **BiblIndex** est un index des citations et allusions bibliques présentes dans la littérature patristique d'Orient et d'Occident. Destiné à faciliter et à renouveler l'étude de la réception des textes scripturaires, BiblIndex peut aussi être utilisé comme une synopse de Bibles en ligne (Outils bibliques), ou comme un index d'œuvres patristiques (Outils patristiques). L'export sera fait en html. Le travail préparatoire a déjà été effectué. Des outils en ligne d'environnements hypertextes (*Hodoi elektronikai, Itinera electronica*) seront utilisés. Il s'agit d'un approfondissement des données existantes. La faisabilité a delà été vérifié à l'aide de corpus échantillon. Les données produites seront au format **.txt**.

Les données produites seront les suivantes :

- Documents textuels des textes antiques non lemmatisés et lemmatisés en langue grec (alphabet grec ancien avec accent, format **Unicode**) en **.txt** regroupés en dossiers par auteurs, c'est-à-dire 6 dossiers de maximum 30 Mo, 30Mo étant le volume de fichiers pour Libanios. A l'heure actuelle, nous disposons de 119 fichiers .txt (16+103) des lemmatisations de Denys d'Halicarnasse (16 doc .txt) et de Libanios (103 doc .txt).
- Documents textuels des passage bibliques liées à l'hospitalité en grec, en latin et en anglais en format .odt car libre, soit 73 documents car il y a 73 livres dans la Bible, fichier individuel d'environ 500Ko
- **Document xml** des résultats Biblindex pour chaque passage étudié avec les références chez les auteurs antiques, soit entre 3000 et 4000 fichiers car la Bible compte 73 livres, pour chaque passage, nous pensons qu'il peut y avoir 50 passages (pour le moment environ 30 pour Genèse et pour Matthieu), 73x50=3650 fichiers.
- **Bibliographie Zotero**, pour gérer les références bibliographiques sur le sujet, maintenue par le porteur de projet
- **Tableur en .csv** contenant la liste des auteurs.

Les données seront accompagnées d'une documentation de type **readme.txt** détaillant la méthodologie utilisée pour collecter les données à partir de Biblindex. Il est possible de créer une **page web didactique** renvoyant vers les différentes ressources pour accompagner les personnes souhaitant réutiliser les données.

**NOMMAGE :** Concernant les conventions de nommage, les noms des livres bibliques sont indiqués en anglais. Le numéro du livre biblique sera séparé du numéro de verset par un underscore et pour les passages, le début et la fin du passage seront séparé par un tiret. Par exemple, Matthew26_7-13 pour Matthieu livre 26 passage du verset 7 au verset 1. Les noms de fichiers ne comporteront pas caractères spéciaux, ni d'espaces.

Chaque nom de fichier se termine par l'indication de la date et de l'heure de la sauvegarde au format américain, soit par exemple 01_13_2020_12 :24, pour le 13 janvier 2020 à 12h24.

Pour les dossiers, le dossier des six auteurs grecs sera séparé du dossier du travail en lien avec la Bible. Le dossier « 6_auteurs_grecs » sera composé de 6 sous-dossiers, correspond à chacun des 6 auteurs, à savoir : Denys d'Halicarnasse, Plutarque, Apollonios de Tyane, Basile de Césarée, Jean Chrysostome et Libanios, soit Dionysos_of_Halicarnassus, Plutarch, Apollonius_of_Tyana, Basil_of_Caesarea, Chrysostomus and Libanios.

**Schéma de la base de données** : la base de données comprendra au minimum deux tables : une table auteur et une table texte. La base de données sera alimentée par des CSV réalisés à la main.

## 2. Documentation et qualité des données

Les données seront accessibles via **ISIDORE**, et seront stockées sur une **HumaNum Box**.

Concernant l'accès aux données, les données seront accessibles dans l'esprit des sciences ouvertes. Les données utilisées dans le cadre de ce projet ne sont pas des données sensibles. Nos données sont des textes antiques libre de droit comme la bible ou le texte des auteurs antiques bruts, n'appartenant pas aux maisons d'éditions. Nous n'avons

pas de données à caractère personnel.

Les références littéraires et les références bibliques seront accessibles librement grâce à la base de données. Les fichiers contenant les textes lemmatisés seront aussi librement accessibles afin de promouvoir l'Open Access. Les données obtenues grâce à l'outil Biblindex répondront aux mêmes principes que le site Biblindex, les données seront accessibles librement tant que le site Biblindex le sera librement. La réutilisation sera conforme à la **licence CC BY NC (Creative Commons),** qui est gratuite et garantie la protection des droits d'auteurs et à la licence Etalab, qui est une licence libre française.

Nous réfléchissons à la question des droits d'auteurs concernant la structure de la base de données, conscients que cette question mérite d'être soulevée.

La qualité des données sera contrôlée grâce à des validations, décidées lors de **réunion de validation** bimensuelle, deux fois par mois. Ces réunions regrouperont le porteur de projet, le post-doctorant en humanités numériques et l'ingénieur d'études. Nous utiliserons des outils permettant des vérifier les liens de la base de données qui renvoient vers les environnements hypertextes, comme par exemple **LinkChecker**, qui est un logiciel libre. Cet outil permet de vérifier s'il n'y a pas de liens cassés dans les documents HTML. **LinkChecker** est un outil python en ligne de commande qui permet de parcourir un site en suivant les liens. Il fournit un résumé (nombre de warning, nombre d'erreurs) et il est configurable pour correspondre à nos besoins. Il sera utilisé par l'ingénieur d'études du projet. Dans le tableur de la base de données, il y aura une **colonne pour indiquer le statut** d'une donnée : « en cours / validé le DATE ». Les données seront soumises à une validation intellectuelle, une validation experte par le porteur du projet.

Concernant la gestion du **versionning**, nous utiliserons un logiciel de gestion des versions : **Git**. Git est un logiciel libre.

## 3. Stockage et sauvegarde pendant le processus de recherche

Les données seront plus facilement interopérables car nous utiliserons les **standards Dublin Core**. Les métadonnées sont moissonnables via le protocole **OAI-PMH**[1].Le protocole OAI-PMH d'interopérabilité des archives ouvertes repose sur un enregistrement minimal **Dublin Core** simple en XML. Les données sont diffusées via Nakala et préservées via l'Huma-Num Box. Huma-Num Box sera utilisée pour l'archivage. The data is disseminated via Nakala and preserved via the Huma-Num Box. Huma-Num Box will be used for archiving.

Les données sont décrites en suivant le standard DublinCore pour garantir l'interopérabilité. Concernant le stockage, les métadonnées sont enregistrées dans Nakala et les données qui sont décrites par lesdites métadonnées sont, stockées dans l'HumaNum Box. Isidore moissonne Nakala et rend accessible ces métadonnées. Nous favoriserons l'utilisation de format ouverts et normalisés. Nous utiliserons des outils pour **contrôler la qualité des formats** de fichier : **l'outil FACILE du CINES** (Centre Informatique National de l'Enseignement Supérieur)[2] et le site du W3C pour la validation des formats xml. Les données seront soumises à une validation technique par l'ingénieur d'études. Nous utiliserons un **vocabulaire contrôlé**, nous utiliserons celui du *Thesaurus Linguae Graecae*.

[1] Open Archives Initiative Protocol for Metadata Harvesting.

[2] Le CINES est un acteur reconnu du domaine de la préservation numérique. FACILE est un outil en ligne de validation de formats.

## 4. Exigences légales et éthiques, codes de conduite

Nos données sont des textes antiques libre de droit comme la bible ou le texte des auteurs antiques bruts, n'appartenant pas aux maisons d'éditions. Nous n'avons pas de données à caractère personnel.

La réutilisation sera conforme à la **licence CC BY NC (Creative Commons),** qui est gratuite et garantie la protection des droits d'auteurs et à la licence Etalab, qui est une licence libre française.

Nous réfléchissons à la question des droits d'auteurs concernant la structure de la base de données, conscients que

cette question mérite d'être soulevée.

/

## 5. Partage des données et conservation à long terme

Les fichiers contenant les textes lemmatisés seront aussi librement accessibles afin de promouvoir l'Open Access. Les données obtenues grâce à l'outil Biblindex répondront aux mêmes principes que le site Biblindex, les données seront accessibles librement tant que le site Biblindex le sera librement.

Le projet pourra bénéficier de la gamme d'outils et de service comme **Nakala** pour l'exposition des données. Le stockage sera délégué à **Humanum** et la préservation sera gérée par **Humanum**. Le but est de pouvoir conserver le document et l'information qu'il contient pour la durée du projet.

Les données pourront être retrouvées et partagées par le dépôt dans un entrepôt de données de confiance, comme **Nakala**.

Les données comme les textes lemmatisées pourront être réutilisé dans le cadre d'autres projets de recherche ou pour l'enseignement.

L'accès aux données sera faisable via **ISIDORE. ISIDORE** consulte les données stockées dans **Nakala**. Nous utiliserons un schéma de métadonnées standard comme le **Dublin Core**, ceci permettra de rendre interopérables les métadonnées, c'est-à-dire la possibilité de pouvoir les connecter à d'autres entrepôts existants, et de les rendre moissonnables par des services spécialisés comme **ISIDORE**.

/

Les données auront un **identifiant unique** : un **Handle Nakala**.

## 6. Responsabilités et ressources en matière de gestion des données

**Bibliographie Zotero**, pour gérer les références bibliographiques sur le sujet, maintenue par le porteur de projet

La qualité des données sera contrôlée grâce à des validations, décidées lors de **réunion de validation** bimensuelle, deux fois par mois. Ces réunions regrouperont le porteur de projet, le post-doctorant en humanités numériques et l'ingénieur d'études.

Les données seront soumises à une validation intellectuelle, une validation experte par le porteur du projet.

Le gestionnaire de données sera l'ingénieur recruté. Il gérera : le dépôt sur **Nakala**, la bibliographie Zotéro et l'évolution du schéma de la base de données. La saisie des données et la production des métadonnées sera assurée par les chercheurs du projet. L'ingénieur s'occupera de la qualité des données, du stockage et de la sauvegarde ainsi que de l'archivage et du partage des données. L'ingénieur sera responsable de la mise en œuvre du plan de gestion de données. La saisie des données et la production des données sera assurée par le post-doctorant. Il sera aussi chargé de l'harmonisation des données. Le **plan de gestion de données sera révisé et mis à jour régulièrement** (avec au minimum 3 versions au cours du projet) lors d'une réunion réunissant le PI, le post-doctorant et l'ingénieur. La personne référente à la fin de l'ERC concernant les gestions des données sera le porteur du projet.

La gestion des données occupera environ 25% du temps de l'ingénieur, soit environ 8h45 par semaine. Les ressources nécessaires pour la diffusion des données sont : les entrepôts de données comme **Nakala**.

Les coûts à prévoir en dehors des coûts de ressources humaines sont par exemple, l’achat de disques durs externe (2 disques de 2To, entre 500€ et 650€ pour un disque dur SanDisk de 2To.