Plan de gestion de données pour le projet e-NdP

Pierre Brochard, Julie Claustre

▶ **To cite this version:**

Pierre Brochard, Julie Claustre. Plan de gestion de données pour le projet e-NdP. 2021. hal-03419065

HAL Id: hal-03419065
https://hal.science/hal-03419065

Submitted on 8 Nov 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution - ShareAlike 4.0 International License

E-NdP

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

Créateurs du PGD : Pierre Brochard, Darwin Smith, Julie Claustre

Affiliation du créateur principal : CNRS

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

Dernière modification du PGD : 08/11/2021

Financeur : Agence nationale de la recherche (ANR)

Numéro de subvention : ANR-20-CE27-0012

Résumé du projet :

Le projet *e*-NDP (Anr 2020) engage l'examen des registres capitulaires, de la documentation et des livres de Notre-Dame de Paris au Moyen Âge, pour étudier la société, l'économie, le bâti du cloître et de ses dépendances. Porté par le LaMOP, le projet a pour partenaires les Archives nationales, la Bibliothèque nationale de France (Département des Manuscrits, Bibliothèque de l'Arsenal), l'École nationale des Chartes et la Bibliothèque Mazarine.

Chercheur Principal : Julie Claustre

Identifiant ORCID : 0000-0001-8504-3920

Contact pour les Données : Pierre Brochard (https://orcid.org/0000-0003-1955-556X)

Produits de recherche :

1. Manuscrits : Numérisations des différents manuscrits (Image)
2. Textes : Transcriptions des différents manuscrits (Texte)
3. Modèles : Modèles de transcription automatique de l'écriture manuscrite (Autre : Modèle IA)
4. Base de données : Base de données comprenant les Entités Nommées (Jeu de données)
5. Cartes : Données Cartographiques (Jeu de données)

E-NdP

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

Manuscrits : Numérisations des différents manuscrits

1a. Comment de nouvelles données seront-elles recueillies ou produites et/ou comment des données préexistantes seront-elles réutilisées ?

Les campagnes de numérisation des manuscrits sont réalisées par les Archives Nationales, par la Bibliothèque Nationale de France et par la Bibliothèque Mazarine.

1b. Quelles données (types, formats et volumes par ex.) seront collectées ou produites ?

Les données produites seront

- des numérisations de 26 registres du Chapitre Notre-Dame de Paris sur la période de 1326 à 1504. Ces numérisations seront au format TIFF, au nombre de 9618 et d'une volumétrie totale de 400 Go. Le format TIFF choisi permet une numérisation de la meilleure qualité possible.
- des numérisations issues du fonds 'Notre-Dame', soit 30 manuscrits à la Bibliothèque nationale de France et 17 manuscrits à la Bibliothèque de l'Arsenal. Ces numérisations seront au format TIFF.

Textes : Transcriptions des différents manuscrits

1a. Comment de nouvelles données seront-elles recueillies ou produites et/ou comment des données préexistantes seront-elles réutilisées ?

Les textes produits sont des transcriptions des manuscrits à l'aide de deux plateformes : eScriptorium (https://escriptorium.fr/) projet de recherche soutenu par PSE et l'Inria et Transkribus (https://transkribus.eu), plateforme payante proposée par la coopérative READ-COOP.

La transcription avec ces deux outils s'effectuent en deux étapes : la première étape consiste à établir une transcription manuelle de plusieurs manuscrits. Dans un deuxième temps, à partir de ces vérités de terrain créées (données d'entraînement), les outils effectuent une transcription supervisée automatique.

L'utilisation de la plateforme eScriptorium sera gratuite pendant la durée de l'ANR.

Le coût d'une transcription effectuée par la plateforme Transkribus à la date du 18/10/2021 serait de 0.2 euros par page (https://readcoop.eu/transkribus/credits/)

1b. Quelles données (types, formats et volumes par ex.) seront collectées ou produites ?

Une transcription pour chaque page de manuscrit sera proposée dans plusieurs formats, un format Texte, un format TEI et un format PageXML.

Le format Texte permettra un usage simple de la transcription effectuée, le format TEI sera destiné à un usage plus élaboré avec l'inclusion de métadonnées décrivant le texte.

Enfin le format PageXML est destiné à un partage de données d'apprentissage pour les outils permettant la reconnaissance de l'écriture automatique. En effet, les données d'apprentissage ainsi créées pourront permettre l'enrichissement d'autres corpus d'apprentissage ou servir à la transcription automatique de textes à la graphie proche.

Modèles : Modèles de transcription automatique de l'écriture manuscrite

1a. Comment de nouvelles données seront-elles recueillies ou produites et/ou comment des données préexistantes seront-elles réutilisées ?

Les modèles seront produits à partir des données d'apprentissage créées dans l'outil eScriptorium sur l'infrastructure CREMMA (Consortium pour la reconnaissance d'écriture manuscrite des matériaux anciens), infrastructure portée par l'École des Chartes et l'Inria.

1b. Quelles données (types, formats et volumes par ex.) seront collectées ou produites ?

Les modèles seront proposés dans un format propre au logiciel Kraken (http://kraken.re/). Ces modèles d'apprentissage pourront servir à la transcription automatique de textes à la graphie proche.

Base de données : Base de données comprenant les Entités Nommées

1a. Comment de nouvelles données seront-elles recueillies ou produites et/ou comment des données préexistantes seront-elles réutilisées ?

Une liste d'Entités Nommées (Lieux et Personnes) sera extraite des transcriptions.
Cette base de données est construite en deux étapes : la première étape consiste à établir une détection manuelle de différentes entités nommées (Personnes, Lieux) afin de créer des données d'apprentissage.
Dans un deuxième temps, à partir de ces données d'entraînement, des librairies de détection d'entités nommées tel la librairie Spacy (https://spacy.io/) effectuent une détection automatique.

1b. Quelles données (types, formats et volumes par ex.) seront collectées ou produites ?

Ces données seront proposées dans un format CSV.
Les Personnes et les Lieux seront alignés dans la mesure du possible avec le référentiel d'autorité Idref (https://www.idref.fr/) et avec création d'un identifiant pérenne de type IdRef pour les entités manquantes.

Cartes : Données Cartographiques

1a. Comment de nouvelles données seront-elles recueillies ou produites et/ou comment des données préexistantes seront-elles réutilisées ?

En croisant les données du cadastre de Vasserot (1810-1836) avec les données des registres capitulaires et d'autres documents du chapitre, il sera possible de reconstituer les données spatiales concernant le bâti et le parcellaire du cloître Notre-Dame avec des degrés de précision variables selon les états chronologiques.

1b. Quelles données (types, formats et volumes par ex.) seront collectées ou produites ?

Les données produites seront de type Shapefile. Elles seront intégrées et distribuées par la plateforme d'information géohistorique sur Paris,

Alpage (https://alpage.huma-num.fr/). La volumétrie des données produites sera précisé ultérieurement.

2. Documentation et qualité des données

2a. Quelles métadonnées et quelle documentation (par exemple méthodologie de collecte et mode d'organisation des données) accompagneront les données ?

Les manuscrits seront disponibles soit sur le site Internet des Archives Nationales, soit sur le site de la BNF, soit sur le site Internet de la Bibliothèque Mazarine. Les manuscrits seront décrits par des métadonnées au format EAD utilisées par ces institutions pour l'ensemble de leurs documents.

Les manuscrits, les textes, les modèles IA et les bases de données seront déposés sur l'entrepôt de données Nakala.

Cet entrepôt de données permet de décrire les données à l'aide du vocabulaire Dublin Core (https://www.dublincore.org), soit le titre de la page, une description, des mots clefs associés, l'ensemble des contributeurs, la date de création et l'historique des dates de modifications du document, un identifiant DOI, etc..). Les liens entre les différents éléments (transcriptions, bases de données, numérisations) seront renseignés par différents champs (dc:source, dc:isreferencedby,): ces métadonnées permettront par exemple à un utilisateur d'effectuer la comparaison entre une transcription d'un manuscrit et ce manuscrit.

Le dépôt sur un entrepôt de données ainsi que l'utilisation d'un vocabulaire contrôlé permettront de répondre aux principes FAIR.

2b. Quelles mesures de contrôle de la qualité des données seront mises en œuvre ?

Le contrôle de la qualité des numérisations des manuscrits sera effectué par les services compétents des Archives Nationales, de la BNF et de la Bibliothèque Mazarine.

Le contrôle de la qualité de la transcription des registres capitulaires sera effectué par un comité éditorial constitué d'experts scientifiques des corpus étudiés.

Les bases de données seront accompagnées par un schéma de données (de type schema.json) décrivant les différents champs ainsi que leurs valeurs possibles. Ces schémas de données permettront de valider le contenu des bases de données. Un historique des modifications des données sera également mis en place.

Le 1er volume du catalogue de 1733 édité au format TEI sera proposé au projet THECAE (Corpus d'inventaires anciens de livres manuscrits et imprimés https://www.unicaen.fr/services/puc/sources/thecae/accueil), cette proposition sera examinée et soumise à l'approbation d'un comité éditorial.

3. Stockage et sauvegarde pendant le processus de recherche

3a. Comment les données et les métadonnées seront-elles stockées et sauvegardées tout au long du processus de recherche ?

Stockage et sauvegarde pendant le processus de recherche :

Les données de recherche (Transcriptions, Modèles, Base de données) produites pendant le processus de recherche seront stockées sur les différents outils de travail proposés par Huma-Num (Gitlab, Sharedocs, Huma-Num Box, Nakala).
Les manuscrits et les données (à l'exclusion de celles déposés sur le dispositif Huma-Num Box) sont sauvegardées par un automate une fois par nuit sur bande magnétique. Une copie de secours est effectuée dans un second bâtiment adjacent au premier.
Les caractéristiques de la sauvegarde journalière sur le dispositif de bandes magnétiques proposés par Huma-Num sont
- en mode incrémental
- par le logiciel TSM (Tivoli) d'IBM
- de tous les fichiers modifiés par rapport à la veille
- avec conservation de 6 versions du même fichier
- avec rétention durant 1 an d'un fichier totalement supprimé à la source
- sur deux robotiques de bandes LTO dans les deux bâtiments distincts

Une sauvegarde de ces données de recherche sera aussi effectué mensuellement sur un serveur de Paris 1 Panthéon-Sorbonne, cette redondance permettant de se prémunir d'un incident grave.
Les données présentes sur le dispositif Huma-Num Box seront copiés sur deux lieux de stockage différents sur disque ainsi que sur bande.

3b. Comment la sécurité des données et la protection des données sensibles seront-elles assurées tout au long du processus de recherche ?

En cas d'incident, les données seront restaurées sur demande auprès des services d'Huma-Num ou depuis le serveur de l'université Paris 1 - Panthéon Sorbonne.

Les données ne comportent pas de données sensibles.

Seuls les membres participant au projet E-NdP pourront accéder en écriture aux données de recherche par l'intermédiaire d'une authentification de type login/motdepasse. L'authentification est géré dans la majorité des outils par l'outil de gestion de login Human-Id (https://humanid.huma-num.fr/)

4. Exigences légales et éthiques, codes de conduite

4a. Si des données à caractère personnel sont traitées, comment le respect des dispositions de la législation sur les données à caractère personnel et sur la sécurité des données sera-t-il assuré ?

Question sans réponse.

4b. Comment les autres questions juridiques, comme la titularité ou les droits de propriété intellectuelle sur les données, seront-elles abordées ? Quelle est la législation applicable en la matière ?

Les données (manuscrits et données) sont diffusées sous Licence Creative Commons Attribution 4.0 (CC BY 4.0) : les données sont librement ré-utilisables, sous condition de citer leurs auteur·es et/ou leurs origines.

A l'exception des manuscrits de la Bibliothèque Mazarine qui sont diffusés sous licence Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International (CC BY-NC-ND 4.0.

4c. Comment les éventuelles questions éthiques seront-elles prises en compte, les codes déontologiques respectés ?

L'ensemble des données n'implique pas de problèmes éthiques.

5. Partage des données et conservation à long terme

5a. Comment et quand les données seront-elles partagées ? Y-a-t-il des restrictions au partage des données ou des raisons de définir un embargo ?

Les données produites seront partagées au fur et à mesure de leur finalisation afin de permettre leurs réutilisations et de favoriser leurs études au delà des personnes impliquées dans l'ANR E-NdP. Les numérisations des manuscrits seront partagées lors la première année du projet.

Les numérisations des manuscrits issues des Archives Nationales seront accessibles sur le site des Archives Nationales par navigation sur le site des Archives Nationales (https://www.archives-nationales.culture.gouv.fr/) et sur l'entrepôt de données d'Huma-Num Nakala par navigation sur le site Nakala (https://www.nakala.fr/) et par le protocole IIIF.

Les numérisations des manuscrits issues de la Bibliothèque Mazarine et de la BNF seront accessibles par le protocole IIIF et par navigation sur le site de la Bibliothèque Mazarine (https://www.bibliotheque-mazarine.fr/fr/) et de la BNF (https://gallica.bnf.fr/)

Les textes, base de données, modèles d'Intelligence Artificielle seront versés sur l'entrepôt de données Nakala.

Un identifiant pérenne de type DOI sera attribué à chaque numérisation, textes, base de données, modèles d'Intelligence Artificielle déposée sur l'entrepôt de données Nakala. Leurs métadonnées décrivant ces données seront moissonnées par les moteurs de recherche usuels (Google, Bing, etc..) mais aussi par le moteur de recherche d'Huma-Num Isidore (https://isidore.science) et par le consortium Datacite (https://search.datacite.org) afin de faciliter leur dissémination.

Cet ensemble de données sera téléchargeable par leur DOI, par l'API d'accès proposé par Nakala (https://api.nakala.fr) et par navigation sur le site de l'entrepôt de données Nakala.

Un identifiant pérenne de type ark sera attribué à chaque numérisation déposée sur le site de la Bibliothèque Mazarine.

Il n'est prévu aucune restriction sur le partage de données.

5b. Comment les données à conserver seront-elles sélectionnées et où seront-elles préservées sur le long terme (par ex. un entrepôt de données ou une archive) ?

Les numérisations des manuscrits seront conservés à la fois par les Archives Nationales et sur l'entrepôt de données Nakala.

Les transcriptions finales des registres, les modèles d'IA ainsi que la base finale des entités nommées seront conservés sur l'entrepôt de données Nakala.

La sauvegarde des données des Archives Nationales et de la Bibliothèque Mazarine est effectuée par le centre national du microfilm et de la numérisation (CNMN) sur le site de St-Gilles-du-Gard.

5c. Quelles méthodes ou quels outils logiciels seront nécessaires pour accéder et utiliser les données ?

L'ensemble des données sera accessible à tout public :

soit à l'aide d'un simple navigateur à travers les sites des Archives Nationales, de la Bibliothèque Mazarine et des services d'exposition des données proposés par Huma-Num (Nakala, Recherche-Isidore, Site Internet du Projet)

soit par l'API de Nakala (https://api.nakala.fr/doc), cet interface permettant aux métadonnées des données du projet d'être machine-actionable.

5d. Comment l'attribution d'un identifiant unique et pérenne (comme le DOI) sera-t-elle assurée pour chaque jeu de données ?

A chaque manuscrit sera attribué soit un permalien (Archives Nationales), soit un ark (Bibliothèque Mazarine) soit un DOI (Nakala).

A chaque texte, modèle d'IA et base de données sera attribué un DOI (Nakala).

6. Responsabilités et ressources en matière de gestion des données

6a. Qui (par exemple rôle, position et institution de rattachement) sera responsable de la gestion des données (c'est-à-dire le gestionnaire des données) ?

Pierre Brochard (https://orcid.org/0000-0003-1955-556X) est le responsable des données issues du traitement des registres capitulaires.

Hélène Noizet (https://orcid.org/0000-0003-2221-6104) est la responsable des données spatiales du projet.

Patrick Latour est le responsable des données issues de la transcription du catalogue de la Bibliothèque de 1733.

6b. Quelles seront les ressources (budget et temps alloués) dédiées à la gestion des données permettant de s'assurer que les données seront FAIR (Facile à trouver, Accessible, Interopérable, Réutilisable) ?

Question sans réponse.