Plans de Gestion de Données des projets membres du consortium CAHIER

Laurene L'Hermite, Fatiha Idmhand, Stéphanie Dord-Crouslé, Karine Abiven, Gaël Lejeune, Alexandre Bartz, Emmanuelle Chapron, Michèle Brunet, Brigitte Gauvin, Thierry Buquet, et al.

Plans de gestion de données

Projets membres du consortium CAHIER

Liste des rédacteurs :

- L'HERMITE, Laurène, IdRef : https://www.idref.fr/236176927 ; Université de La Rochelle, Centre de recherches en histoire internationale et atlantique (EA1163), La Rochelle, France
- IDMHAND, Fatiha, IdHAL : fatiha-idmhand ; ORCID : 0000-0001-7135-9182 , Université de Poitiers, Institut des textes et manuscrits Modernes, ITEM (UMR-CNRS 8132), Poitiers, France
- ABIVEN, Karine, IdHAL : karine-abiven ; ORCID : https://orcid.org/0000-0001-9518-1040, Sorbonne Université, Sens-Texte-Informatique-Histoire (STIH), UR 4509, France
- LEJEUNE, Gaël, IdHAL : gael-lejeune ; ORCID : https://orcid.org/0000-0002-4795-2362, Sorbonne Université, Sens-Texte-Informatique-Histoire (STIH), UR 4509, France
- BARTZ, Alexandre. IdHAL : alexandre-bartz ; ORCID : https://orcid.org/0000-0003-0850-8266, Sorbonne Université, Sens-Texte-Informatique-Histoire (STIH), UR 4509, France
- CHAPRON, Emmanuelle, IdHAL : emmanuelle-chapron ; ORCID : https://orcid.org/0000-0001-9907-7961, Aix-Marseille Université, CNRS - TELEMMe (UMR 7303), Marseille, France
- BRUNET, Michèle, IdHAL : michele-brunet ; ORCID : https://orcid.org/0000-0003-1818-5237, Université Lyon 2, Histoire et Sources des Mondes Antiques (HiSoMA, UMR 5189), Lyon, France
- GAUVIN, Brigitte, ISNI : 0000000061551536, Université de Caen-Normandie, CRAHAM (UMR 6273), Caen, France
- BUQUET, Thierry, IdHal : thierry-buquet, Orcid : 0000-0003-2956-8217 ; CNRS, CRAHAM (UMR 6273), Caen, France
- BUARD, Pierre-Yves, Pôle Document numérique, MRSH, Université de Caen Normandie – CNRS (USR 3486), Caen, France
- ANDRISI-BRÉMON, Cécile, IdHAL : 1084562 ; IdRef : 193185725 ; ORCID : https://orcid.org/0000-0002-5130-339X, Autoentrepreneuse pour le compte de l'Université de Paris, en lien avec le CERILAC (UPR n°441), Paris, France ; ingénieure d'études en CDD d'octobre 2015 à juin 2019 (Centre Seebacher, CERILAC, Université Paris Diderot)
- RITZ, Olivier, IdHAL : olivier-ritz ; ORCID : 0000-0001-5492-9403, Université de Paris, CERILAC (UPR 441), Paris, France
- PETITIER, Paule, IdHAL : 720408 ; ISNI : 0000000117567255 ; IdRef : 033175756 , Université de Paris, CERILAC (UPR 441), Paris, France
- DORD-CROUSLE, Stéphanie, IdHAL : stephanie-dord-crousle ; ORCID : 0000-0002-6683-9509, CNRS (UMR 5317 IHRIM), France
- NOILLE, Christine, ISNI : 0000000109073343 ; IdRef : https://www.idref.fr/032141025 professeure, Sorbonne Université, UMR 8599 CELLF, France

Table des matières

Antonomaz (« ANalyse auTOmatique et NumérisatiOn des MAZarinades »)

1. Plan de gestion de données (PGD) du projet Antonomaz (« ANalyse auTOmatique et NumérisatiOn des MAZarinades »)

- ***Présentation de la section***

Cette section décrit le PGD : elle présente l'auteur du PGD, les relecteurs du PGD, les autres intervenants assurant la gestion du PGD et, le cas échéant, ses mises à jour.

- ***Recommandations :***

Il est utile de désigner un responsable du PGD qui sera la personne à contacter. Il n'est pas nécessairement le responsable scientifique du projet. Il est recommandé d'associer ce responsable à son identifiant ORCID, IdRef, ISNI, IdHal et de nommer l'ensemble des personnes ayant contribué à la rédaction et à la relecture du PGD.

Le PGD évolue au fur et à mesure de l'avancée du projet de recherche et de l'enrichissement des données. Afin de faciliter sa rédaction, il est conseillé d'en produire une première version au début du projet, qui sera modifiée éventuellement en cours de projet, ainsi qu'à la fin du projet et d'indiquer les versions du PGD dans leur ordre antéchronologique en commençant par l'actuelle.

Auteurs du plan de gestion des données :

ABIVEN, Karine, IdHAL : karine-abiven ; ORCID : *https://orcid.org/0000-0001-9518-1040*, Sorbonne Université, Sens-Texte-Informatique-Histoire (STIH), UR 4509, France
Rôle dans le projet : Responsable du projet

LEJEUNE, Gaël, IdHAL : gael-lejeune ; ORCID : https://orcid.org/0000-0002-4795-2362, Sorbonne Université, Sens-Texte-Informatique-Histoire (STIH), UR 4509, France
Rôle dans le projet : Co-responsable du projet

BARTZ, Alexandre. IdHAL : alexandre-bartz ; ORCID : https://orcid.org/0000-0003-0850-8266, Sorbonne Université, Sens-Texte-Informatique-Histoire (STIH), UR 4509, France
Rôle dans le projet : Ingénieur du projet (en 2021).

L'HERMITE, Laurène, IdRef : https://www.idref.fr/236176927 ; Université de La Rochelle, Centre de recherches en histoire internationale et atlantique (EA1163), La Rochelle, France
Rôle dans le projet : co-auteure du PGD

Version du plan de gestion des données :

PGD V1 : 30/10/2021, PGD projet Antonomaz

Deux versions de ce PGD sont actuellement prévues

2. Présentation du projet et responsabilités

- ***Présentation de la section***

Cette section décrit le projet ou le corpus sur lequel porte le PGD. Elle décrit le projet, ses objectifs, participants, etc. Ici, nous décrivons le Consortium CAHIER mais vous trouverez en annexe des exemples plus précis basés sur des projets membres de CAHIER

- ***Recommandations***

Si le nom du projet est un acronyme, indiquez également la version développée.
Exemple : Antonomaz ("ANalyse auTOmatique et NumérisatiOn des MAZarinades")
Identifier la/le responsable scientifique du projet : Nom, Prénom, Institution, Laboratoire, Unité de rattachement, Ville, Pays. Mettre en lien son identifiant ORCID (ou ISNI, IdRef, IdHal, ...). Si possible indiquez des données de contact (courriel, téléphone professionnel)
Exemple : Karine ABIVEN (https://orcid.org/0000-0001-9518-1040), Sens-Texte-Informatique-Histoire (STIH), EA 4509, Université Paris - Sorbonne, Paris IV, France
Précisez également si le projet s'inscrit dans une programmation scientifique financée et les axes scientifiques liés à cette programmation :
Axe scientifique d'un Labex
Programme de financement d'un projet ANR, H2020
Axe ou programme scientifique d'une structure de recherche liée au porteur ou à l'équipe projet...

Nom du projet

Antonomaz ("ANalyse auTOmatique et NumérisatiOn des MAZarinades")

Responsable du projet (principal researcher) et unité de rattachement

ABIVEN, Karine, IdHAL : karine-abiven ; ORCID : *https://orcid.org/0000-0001-9518-1040*, Sorbonne Université, Sens-Texte-Informatique-Histoire (STIH), UR 4509, France
Rôle dans le projet : Responsable du projet

Financeur(s) du projet et type de financement

DIM STCN (Ile de France) en 2019-2021.

IUF (Institut Universitaire de France) en 2020-2025.

Institution / organisme / unité porteuses du projet

Projet dans le cadre de l'Institut Universitaire de France

Partenaires (identifier les organismes partenaires, ressources et co-financeurs du projet)

En partenariat avec la Bibliothèque Mazarine, et cofinancé par l'OBVIL (SU), le DIM STCN, le consortium CORLI.

Le projet a été labellisé par le Consortium CAHIER [https://cahier.hypotheses.org/membres] de la Tigr Huma-num (voir aussi [https://cahier.hypotheses.org/antonomaz]).

Descriptif et objectif(s) du projet

Le projet Antonomaz, "ANalyse auTOmatique et NumérisatiOn des MAZarinades" vise à exploiter une collection numérique d'environ 5000 écrits ayant pour objet les affaires politiques de la régence du cardinal Mazarin, et traditionnellement appelés "mazarinades".

Notre approche se situe dans le champ des études littéraires, de l'analyse du discours et des Humanités Numériques. Elle vise à fournir des méthodes automatiques, empruntant au Traitement Automatique des Langues, à la fouille de données et à la linguistique de corpus, pour l'analyse de ces données par les spécialistes des divers domaines concernés (littéraires, linguistes, historiens, principalement).

Objectifs :

1/ Offrir une collection numérique de ces écrits, la plus exhaustive possible, en libre accès. Dans ce but, le projet abonde une bibliothèque numérique, celle de la Bibliothèque Mazarine (Mazarinum), en favorisant les numérisations puis en automatisant le passage du mode image au mode texte. Il s'agit d'offrir aux utilisateurs une collection téléchargeable pour des traitements avancés (statistiques, par exemple), pour compléter des ressources comme le signalement en ligne de collections entières (comme celle de l'Université de Tokyo, mise en ligne par le groupe des Recherches internationales sur les Mazarinades). Le but serait aussi de permettre la sélection de corpus cohérents à l'intérieur de cette nouvelle collection numérique, en fonction des besoins de chaque usager.

2/ Améliorer les données textuelles obtenues par des transcriptions automatiques (par reconnaissance de caractères), en mettant à profit les méthodes d'apprentissage profond. Il s'agit de paramétrer finement la reconnaissance automatique des caractères originaux figurant dans l'imprimé ancien.

Rendre disponible ce modèle entraîné, qui pourrait servir à améliorer les sorties en mode « texte » de fac-simile d'autres projets numériques s'intéressant aux "imprimés non-livres" (*non-book printed material*), par exemple la Newberry French Pamphlet Collection

(https://archive.org/details/newberryfrenchpamphlets?and[]=mediatype%3A%22texts%22&and[]=year%3A%221649%22)

3/ Expérimenter des applications en Traitement Automatique des Langues : datation automatique, attribution d'auteur, classification non-supervisée. Ces expériences exploitent d'abord directement les données brutes (sorties d'OCR bruitées), dont l'analyse au grain caractère peut produire des résultats parfois meilleurs que les données lissées pour l'œil humain, bien plus coûteuses pourtant à obtenir.

S'ensuivent plusieurs pistes de travail, comme l'automatisation de la normalisation et l'annotation du mode texte, ainsi que divers types de balisage et d'extractions d'informations comme les entités nommées.

4/ Proposer une visualisation originale des liens entre ces textes polémiques : en raison de leur nature réactionnelle, ils n'ont de sens que pris dans leur contexte fin et compris dans leur mise en réseau. Il s'agira donc de donner un accès visuel dynamique à leur enchaînement à la fois chronologique et réticulaire. Des visualisations des métadonnées seront aussi proposées (par date, par épisodes historiques, par taille de l'imprimé, par éditeurs connus, etc.)

Dates et durée

Date de début de financement et de début des travaux : 2019.

Date de fin de financement et de fin des travaux : 2025.

Mots clés du projet

- Mazarinades
- Langue et littérature françaises
- Traitement Automatique du langage naturel
- Analyse du discours
- Exploration de données
- Data visualisation
- Reconnaissance optique des caractères
- Métadonnées
- Philologie numérique
- Pamphlets

Publications (articles, pré-proposition, site web, …)

Sites web du projet :

Site web à venir. Présentation de la collection numérique de mazarinades : accès aux PDF accessibles en ligne mis en réseau. Possibilité de recherches plein texte, de recherches d'entités nommées, et de recherches textométriques. Visualisations des métadonnées.

Sites de présentation actifs :

- Site du Domaine d'Intérêt Majeur (DIM) Sciences du Texte et Connaissances Nouvelles : http://www.dim-humanites-numeriques.fr/projets/antonomaz/
- Carnet de recherche en lien avec le projet : https://libelles.hypotheses.org/

Bases de données interrogeables en ligne :

- Mise en ligne des numérisations en format IIIF d'un ensemble de libelles conservés à la Mazarine (1ere vague de 419 libelles dont le financement a été assuré par l'OBVIL pour Antonomaz) : https://mazarinades.bibliotheque-mazarine.fr/
- Mise en ligne de la bibliographie Moreau ("Moreau En Ligne"), et de ses suppléments, structurée en format texte et interrogeable par divers champ (numéro Moreau, date, mots de la notice Moreau, nombre de pages, etc.). Une dernière version à jour de Juillet 2021 est en ligne : http ://memes.sorbonne-universite.fr/visualisation/Moreau/test.html

Listes des articles publiés par le projet :

- 2021a : Karine Abiven, Gaël Lejeune, "Des données au corpus : l'exploitation numérique des mazarinades", *Dix ans de Corpus d'auteurs*, Editions des Archives contemporaines, accepté.

- 2021b : **Karine Abiven, Jean-Baptiste Tanguy et Gaël Lejeune**, "Exploiter en corpus des données textuelles océrisées : l'écriture burlesque de la Fronde (1648-1652)", accepté, *revue Humanités numériques*, n°4 - Humanistica.

- 28/06/20 : **Jean-Baptiste Tanguy,** "Exploiter des modèles de langue pour évaluer des sorties de logiciels d'OCR pour des documents français du XVIIe siècle", article accepté à *RECITAL*@TALN 2020,

- 10/03/20 : **Anaëlle Baledent (GREYC, Normandie Université), Nicolas Hiebel et Gaël Lejeune**, "Dating Ancient texts: an Approach for Noisy French Documents", article accepté à *Language Technologies for Historical and Ancient Languages (LT4HALA)*,

- 2019 : **K. Abiven** : « Le moment discursif des barricades d'août 1648 : quelle interprétation des récurrences dans le discours sur l'événement ? », *Cahiers de Narratologie* [En ligne], 35 | 2019, mis en ligne le 03 septembre 2019, URL : http://journals.openedition.org/narratologie/9264

- 29/11/19 : **K. Abiven**: « La liste de noms propres dans les libelles de la Fronde : les revendications de prestige et leur satire », *Journées d'étude Listes de noms. Ordre social et ordre du livre*, M. Roussillon et C. Schuwey, Université d'Artois, Arras.

- 05/04/19 : Séminaire à l'OBVIL : analyse stylistique de textes littéraires

- 21/03/19 : Projet Antonomaz, Séminaire LCSU

- **A. Baledent et G. Lejeune**, "Automatic Stylistic Analysis; a search for efficient and interpretable descriptors to characterize individual writing style", in *Phraséologie et stylistique de la langue littéraire*, Ludwig Fesenmeier et Iva Novakova (eds.), Peter Lang, 2020, p. 329-342.

- 14/03/19 : Anaëlle Baledent et Gaël Lejeune, "Analyse stylistique automatique : A la recherche d'indices efficaces et pertinents pour caractériser le style de Dumas", *Phraseorom* 2019.

- 15/01/19 : **Karine Abiven et Gaël Lejeune**, "Analyse automatique de documents anciens : tirer parti d'un corpus incomplet, hétérogène et bruité", revue *RIDOWS* – Pdf

Autres livrables (guides, recommandations, etc.) :

Un modèle d'OCR entrainé, spécifique aux imprimés de type brochure de l'Ancien Régime : reprise du modèle entraîné par Simon Gabay et Claire Jahan (spécifique aux imprimés du XVIIe s.). Il est prévu d'entraîner ce modèle sur les futures numérisations livrées par la bibliothèque Mazarine (lot 2 des mazarinades). Cet entraînement sera effectué avec E-scriptorium ce qui permet l'utilisation du standard IIIF (en lien avec Biblissima) pour la récupération des images et des métadonnées. Une fois ce modèle entraîné, il sera mis à la disposition de la communauté : diffusion du modèle en licence CC-BY et sur la plateforme GitHub HTR-United.

3. Présentation et description du corpus

- ***Présentation de la section***

Cette section décrit le corpus et ses données. Elle décrit de façon plus précise les données du projet, les méthodes appliquées pour les collecter, etc. Ici, nous décrivons les données du Consortium CAHIER dans leur ensemble mais vous trouverez en annexe des exemples plus précis basés sur des projets membres de CAHIER

- ***Recommandations***

Il s'agira de préciser le mode de collecte et l'origine des données, les centres d'archives, bibliothèques ou centres d'études hébergeant les données y compris si les données procèdent d'un moissonnage de ressources en ligne. L'organisation du corpus, l'arborescence des fichiers, le système de nommage et de gestion des répertoires et des fichiers doit être décrite. De même que la nature des données, leurs formats, leur volumétrie (en poids et nombre de fichiers), leur état, etc. Pour que les données soient réutilisables sur le long terme, les formats doivent être ouverts et non propriétaires et les données stockées dans des entrepôts accessibles.

Nom du projet

Antonomaz, "ANalyse auTOmatique et NumérisatiOn des MAZarinades"

Présenter et décrivez le corpus

L'ensemble recouvre théoriquement 5063 titres connus d'écrits généralement brefs (brochures et libelles) appelés "mazarinades" (sur environ 25 000 exemplaires référencés par la bibliothèque Mazarine, car plusieurs exemplaires d'une même édition ont été conservés). Ces écrits ont été publiés pendant la Fronde (1648-1653). Il s'agit de documents imprimés, dont la teneur est soit en faveur ou en défaveur de la politique du cardinal Mazarin. Cette collection est aussi diversifiée par les genres (chansons, relations, proclamations, satires, etc.) que par la forme des publications (placard, livret, acte officiel, etc.). Le référentiel bibliographique ainsi constitué se fonde notamment sur une bibliographie établie par Célestin Moreau en 1850-1851, outil bibliographique numérisé par le projet, et sa rénovation actuelle par la bibliothèque Mazarine (en janvier 2020, 20% des notices attendues à terme).

Un des enjeux d'Antonomaz est de permettre à l'utilisateur de sélectionner, à l'intérieur de cette collection numérique, des corpus cohérents (par auteur, par genre, par événement, etc. Par exemple : écrits de Scarron, écrits burlesques, écrits relatifs aux barricades de Paris).

Période couverte par le corpus, auteur(s) concerné(s)

1648-1653

Organisation du corpus

Nommage des documents : identifiant Moreau et ses suppléments (avec le même codage que la Base Bibliographique des Mazarinades) et indication de la source numérique. Par exemple Moreau50_GALL (pour un imprimé référencé au numéro 50 par Moreau et trouvé dans la bibliothèque numérique de la BNF, Gallica).

Architecture des Xml sur le Github

Structuration en collections, sous-collections, réseaux de textes : à venir.

Mode de collecte et origine des données

Fac-simile numériques issus des bibliothèques numériques Gallica, Mazarinum, Gbooks. Pour Gallica, collecte semi-automatisée grâce à l'API de cette bibliothèque.

Pour les métadonnées, Moreau en ligne et données bibliographiques issus de la base bibliographique de la Bibliothèque Mazarine.

Etat du corpus numérique

Corpus en cours de production : actuellement (octobre 2021) 2 970 documents en PDF recueillis en ligne, dont 2569 éditions uniques, c'est-à-dire sans compter les différents exemplaires d'une même édition. La collecte a été faite sur les bibliothèques numériques Gallica, Mazarinum, Gbooks. Choix des PDF selon la qualité (dans l'ordre de préférence Mazarinum - de très haute qualité -, Gallica - souvent numérisés sur microfilms -, GBooks - souvent la dernière page est réduplicuée de nombreuses fois).

Une partie des fichiers ont été océrisés à ce jour via une chaîne de traitement adaptée aux textes imprimés du 17e siècle.

Leur passage semi-automatique en xml (2000 documents à ce jour, avec structuration minimale) est en cours ; et les relectures (transcription et relectures du header). Cf volumétrie ci-dessous.

Types de données :

- Images et textes en PDF
- Transcriptions encodées en XML-TEI (avec métadonnées)
- Images IIIF

Volumétrie

Evolution de la volumétrie (en nombre) des données au 27/09/2021

Date	PDF #docs	PDF #pages	XML #docs	XML #pages	XML #tokens	XML #types)	Retranscrits #docs
02/06/2021	1.111	15.000	447	1.500	2.108.211	199.374	105
02/07/2021	2.221	*71.069*	687	*10.423*	2.647.056	242.418	105
22/07/2021	2.613	*80.524*	750	*11.368*	2.811.341	257.506	105
27/09/2021	2.613	*80.524*	*2000*	*30. 000*	*7.350.000*	257.506	105

Modifications effectuées sur les données, versions, …

Océrisation et encodage XML/TEI des imprimés. Versionnage en fonction de l'avancée des relectures.

Autres données créées ou collectées pour documenter et/ou enrichir les corpus constitués.

Liens vers les différents *githubs* créés pour les besoins du projet :

- Outils développés, antonomaz_tools : https://github.com/rundimeco/antonomaz_tools
- Encodage du corpus, Antonomaz : https://github.com/Antonomaz

Métadonnées créées et standards et formats utilisés

Pour les données interrogeables via la base Moreau-En-Ligne :

Les métadonnées utilisées pour le moment (2019-2021) sont celles léguées par la tradition bibliographique, issues du répertoire Moreau. Cet outil en ligne permet de requêter les imprimés par numéro Moreau, par une séquence de caractères du titre, l'année, la date plus précise quand elle est accessible, le lieu de publication, le nombre de pages et la notice Moreau.

Peu à peu, ces métadonnées sont remplacées, notamment dans l'encodage xml des documents, ces métadonnées sont corrigées et enrichies par celles issues des archives d'H. Carrier, qui préparait une bibliographie critique des Mazarinades.
Voir : https://mazarinum.bibliotheque-mazarine.fr/expositions-virtuelles/item/17787-vii-enqueter-sur-les-mazarinades?oeuvre=19#page=1&viewer=picture&o=no&n=0&q=

Ces métadonnées rénovées sont le fruit du travail de la Bibliothèque Mazarine, dans leur "Base Bibliographique des Mazarinades", outil évolutif lancé en 2019 : lhttps://mazarinades.bibliotheque-mazarine.fr/

Les métadonnées descriptives, administratives et techniques

L'en-tête des fichiers XML/TEI contient les informations habituellement requises :

- Les informations sur le texte (titre) et quand disponibles : la date (présente à 98% sur les imprimés), le lieu (environ 80 % continent leur lieu d'impression, avec environ 40 fausses adresses), l'éditeur, l'auteur (mais 80 % d'anonymes), nombre de pages, format (in quarto pour plus de 90% des documents, très homogènes de ce point de vue).

- On renvoie quand disponible, à la notice de la Base Bibliographique de la Mazarine, qui fournit les métadonnées les plus fiables (sur 20% du corpus). Pour les autres, on signale la source de connaissance des métadonnées (bibliographie Moreau ou catalogues de bibliothèque).

- Les noms et les lieux sont liés au web sémantique par des balises contenant l'isni, le geoname et l'identifiant wikidata.

- La balise <MsDesc> inscrit la cote du document physique, le lieu de conservation. On ajoute une balise sur la présence ou non de tampons (avec réponse boléenne) qui peut renseigner sur l'origine du document. A terme on aimerait renseigner la présence ou nom du bref imprimé dans un recueil car c'est une information très importante pour connaitre les usages qui en ont été faits.

- Les mots clés concernent la forme (vers/prose), le ou les genres et sous-genres, le sujet quand récupérable sur la base bibliographique de la Mazaraine, ou quand il a été possible de le renseigner à la main.

- La balise encodingDesc > est ainsi renseignées : «<p> Cette édition a été réalisée dans le cadre du projet ANTONOMAZ. Son objectif principal est de fournir un texte destiné à l'exploration

avec des outils électroniques. De ce fait, ce n'est ni une édition philologique, ni une édition pédagogique ou de redécouverte d'un auteur oublié. »

- Les textes encodés dans le cadre du projet ANTONOMAZ sont issus de numérisations de bibliothèques numériques publiques et de Google livres.

- L'édition présentée ici est issue d'un processus d'OCRisation réalisé avec Kraken.

Les métadonnées structurelles et l'annotation sémantique

Le Format Json a été utilisé pour structurer la liste des titres et la liste des épisodes historiques (qui seront ensuite encodés semi-automatiquement dans le fichier xml).

Les balises TEI servent à la structuration du texte et des métadonnées. Elles sont contrôlées et encadrées par un schéma de validation (ODD) auquel nous renvoyons pour documenter les métadonnées et les principes d'annotation :
Projet Antonomaz, ODD, 2021, consulté le 20/10/2021, URL : https://github.com/Antonomaz/ODD

Référentiels d'indexation utilisés (vocabulaires contrôlés - thésaurus ou ontologies disciplinaires - et/ou indexation libre)

Dans le balisage TEI :
- isni
- geonames
- wikidata
- identifiant unique et URL pérenne pour chaque édition de texte, appelé numéro BM (quand disponible), depuis la Base bibliographique des mazarinades (https://mazarinades.bibliotheque-mazarine.fr/)

4. Modalités de partage, de sauvegarde et de protection des données. Volumétrie des données stockées et espaces choisis.

- ***Présentation de la section***

Cette section décrit la documentation produite au cours projet. Il s'agit d'une documentation autre que numérique (sur support papier par exemple). Si elle existe, il est important de la décrire. Cette section décrit également les lieux et infrastructures de stockage des données pendant le projet.

- ***Recommandations***

Il s'agira de préciser ici le matériel physique et les lieux de stockage des données. Idéalement, il faudrait stocker les données dans au moins 2 endroits, éviter le stockage externe et privilégier les outils mis à disposition par l'institution. Pour cela, il peut être nécessaire de savoir quel est le volume approximatif des données à sauvegarder, l'espace de stockage nécessaire, la périodicité des sauvegardes, le nom et la nature du service fourni par l'institution, etc. On peut également indiquer les procédures de sauvegarde mises en place (fréquence des sauvegardes, automatisée ou non ?), les personnes en charge de la protection de ces données et du contrôle de l'accès, le mode de récupération des données en cas d'incident…

Un serveur avec une page web : https://antonomaz.huma-num.fr
L'accès est ouvert depuis le 15 octobre 2021. Il rassemblera les divers espaces de stockage utilisé jusqu'ici :

- Un wiki destiné à documenter le projet et les données : http://stih-sorbonne-universite.fr/dokuwiki/doku.php?id=antonomaz
- Données d'encodage stockées sur un GitHub : https://github.com/orgs/Antonomaz/repositories
- Une ressource bibliographique interrogeable : http://memes.sorbonne-universite.fr/visualisation/Moreau/test.html
- Ainsi que Sharedocs, espace de partage d'Huma-num, utilisé en interne pour stocker les données.

Accès, partage et limites des données

Données primaires publiques.
Fichiers en format texte produits et accessibles sur le github du projet : Licence Creative Commons CC-by.
Métadonnées de la « Bibliographie des Mazarinades » : Licence Creative Commons CC-by-nc-nd.

Sources en format PDF (fac-simile numériques) :

Gallica : Licence ODBL.
Mazarinum: Licence CC-by-nc-nd
GBooks : ne pas utiliser les fichiers à des fins commerciales, ne pas supprimer l'attribution Google.

5. Responsabilités et ressources pour la gestion des données

- ***Présentation de la section***

Cette section décrit, identifie, présente et nomme les responsables de la gestion des données.

- ***Recommandations***

Afin de respecter les principes FAIR, CAHIER recommande le dépôt de celles-ci dans l'entrepôt Nakala (https://www.nakala.fr/). Ce service de dépôt et de stockage des données est proposé par la TGIR HumaNum pour les SHS. Il assure la gestion pérenne et sûre des données. Utiliser Nakala n'empêche pas de recourir à un second dépôt sur un autre entrepôt ou sur une plateforme institutionnelle.

Responsables de la gestion des données :

ABIVEN, Karine, IdHAL : karine-abiven; ORCID : https://orcid.org/0000-0001-9518-1040, Université Sorbonne,Sens-Texte-Informatique-Histoire (STIH), UR 4509, France

LEJEUNE, Gaël, IdHAL : gael-lejeune ; ORCID : https://orcid.org/0000-0002-4795-2362, Sorbonne Université, Sens-Texte-Informatique-Histoire (STIH), UR 4509, France

Évaluation des coûts (budgets, personnels et temps) dédiés à rendre les données FAIR (temps et budgets pour la collecte et la diffusion des données, pour le stockage et l'archivage).

Equipe engagée dans la gestion des données à différentes étapes du projet :

Actuels :

- Karine Abiven MCF en Langue Française
- Gaël Lejeune MCF en Informatique
- Jean-Baptiste Tanguy Doctorant en Humanités Numériques
- Alexandre Bartz, Ingénieur

Anciens :

- 2021 : Mélanie Lecha, M2 Humanités Numériques, ENS Lyon/ENSSIB
- 2021 : Camille Roblin, M2 Humanités Numériques, Lyon 2/ENSSIB
- 2021 : Amélie Hip, vacataire (retranscriptions)
- 2019-2020 : Sylia Kecili, stagiaire (M1 TILDE, Paris 13) sur les problématiques d'OCR
- 2018-2019 :
    - Anaëlle Baledent, stagiaire (M2) sur la datation, actuellement en thèse d'Informatique à l'Université de Caen
    - Nicolas Hiebel, stagiaire (L3) sur la datation, actuellement en M2 Langue et Informatique à Sorbonne Université
    - Jamiilah Patel, stagiaire (M1), sur la structuration des métadonnées, Masterante en Littérature à Sorbonne Université

Évaluation des coûts : 203.5 KE

- Masse salariale, via le DIM STCN et l'OBVIL (2 stagiaires, 1 vacataire non étudiante, 1 IGE) : 61 KE

- Postes de travail pour doctorant et stagiaire (CORLI, faculté des lettres Sorbonne Université) : 5 KE
- Numérisation de 800 documents avec les standards des bibliothèques publiques (IUF) : 16 KE
- Thèse (financée par la région via le DIM STCN) : 100 KE
- Contribution de la Bibliothèque Mazarine estimée à 21 KE (en Personnel : Mise en œuvre et suivi de la numérisation des sources, production et structuration des métadonnées, contrôle qualité de la numérisation ; Elaboration et maintenance de la Bibliographie des mazarinades [BM], référentiel en ligne dont l'ouverture est programmée pour juin 2019. En matériel : mise à disposition du matériel de numérisation)
- Conférences (Bordeaux 2020 : financé par Cahier, Orléans 2021 par l'IUF) : 500 E

6. Archivage des données

- ***Présentation de la section***

Cette section décrit les données à conserver à court, moyen et long terme, les éventuelles données à détruire ou à laisser sous embargo et indique la durée de cette restriction.

- ***Recommandations***

A l'issue du projet, des jeux de données se prêteront à une conservation à long terme pour une utilisation future, tandis que d'autres données ne nécessitent qu'une préservation à moyen terme car jugées moins essentielles et au potentiel de réutilisation limité, voire, elles pourront être destructibles pour des raisons de légalité ou de confidentialité.

Plateforme pour l'archivage pérenne des données

Zenodo pour les fichiers XML (qui applique les principes FAIR)
ou CINES : https://www.cines.fr/archivage/ (mais peut-être surdimensionné).

Durée de conservation des données

Illimité

Volume des données à conserver

Volume peu important (<5 Go si l'on ne compte que les fichiers XML-TEI : version actuelle et version lemmatisée).

Les documents (fac-similes numériques) sont hébergés sur les plateformes des bibliothèques numériques et non par le projet.

Coûts alloués à la conservation

À voir en fonction de la plateforme choisie.

Outils, méthodes, procédures nécessaires pour accéder à ces données archivées et les réutiliser

Données en accès ouvert (pour les limitations, voir les licences des bibliothèques numériques ci-dessus)

Archives savantes des Lumières. Correspondance, collections et papiers de travail d'un savant nîmois : Jean-François Séguier (1703-1784)

1. Plan de gestion de données (PGD) du projet d'édition de la correspondance de Jean-François Séguier

- ***Présentation de la section***

Cette section décrit le PGD : elle présente l'auteur du PGD, les relecteurs du PGD, les autres intervenants assurant la gestion du PGD et, le cas échéant, ses mises à jour.

- ***Recommandations***

Il est utile de désigner un responsable du PGD qui sera la personne à contacter. Il n'est pas nécessairement le responsable scientifique du projet. Il est recommandé d'associer ce responsable à son identifiant ORCID, IdRef, ISNI, IdHal et de nommer l'ensemble des personnes ayant contribué à la rédaction et à la relecture du PGD.

Le PGD évolue au fur et à mesure de l'avancée du projet de recherche et de l'enrichissement des données. Afin de faciliter sa rédaction, il est conseillé d'en produire une première version au début du projet, qui sera modifiée éventuellement en cours de projet, ainsi qu'à la fin du projet et d'indiquer les versions du PGD dans leur ordre antéchronologique en commençant par l'actuelle.

Auteures du plan de gestion des données :

CHAPRON, Emmanuelle, IdHAL : emmanuelle-chapron ; ORCID : https://orcid.org/0000-0001-9907-7961, Aix-Marseille Université, CNRS - TELEMMe (UMR 7303), Marseille, France

L'HERMITE, Laurène, IdRef : https://www.idref.fr/236176927 ; Université de La Rochelle, Centre de recherches en histoire internationale et atlantique (EA1163), La Rochelle, France
Rôle dans le projet : co-auteure du PGD

Version du plan de gestion des données :

PGD V1 : 30/10/2021
PGD projet Archives savantes des Lumières. Correspondance, collections et papiers de travail d'un savant nîmois : Jean-François Séguier (1703-1784)
1 version de ce PGD est actuellement prévue

2. Présentation du projet et responsabilités

- ***Présentation de la section***

Cette section décrit le projet ou le corpus sur lequel porte le PGD. Elle décrit le projet, ses objectifs, participants, etc. Ici, nous décrivons le Consortium CAHIER mais vous trouverez en annexe des exemples plus précis basés sur des projets membres de CAHIER

- ***Recommandations***

Si le nom du projet est un acronyme, indiquez également la version développée.
Exemple : Antonomaz ("ANalyse auTOmatique et NumérisatiOn des MAZarinades")
Identifier la/le responsable scientifique du projet : Nom, Prénom, Institution, Laboratoire, Unité de rattachement, Ville, Pays. Mettre en lien son identifiant ORCID (ou ISNI, IdRef, IdHal, ...). Si possible indiquez des données de contact (courriel, téléphone professionnel)
Exemple : Karine ABIVEN (https://orcid.org/0000-0001-9518-1040), Sens-Texte-Informatique-Histoire (STIH), EA 4509, Université Paris - Sorbonne, Paris IV, France
Précisez également si le projet s'inscrit dans une programmation scientifique financée et les axes scientifiques liés à cette programmation :
Axe scientifique d'un Labex
Programme de financement d'un projet ANR, H2020
Axe ou programme scientifique d'une structure de recherche liée au porteur ou à l'équipe projet...

Nom du projet

Écritures savantes au siècle des Lumières. La correspondance et les carnets de visiteurs de Jean-François Séguier

Responsable du projet (principal researcher) et unité de rattachement

CHAPRON, Emmanuelle, IdHAL : emmanuelle-chapron ; ORCID : https://orcid.org/0000-0001-9907-7961, Aix-Marseille Université, CNRS - TELEMMe (UMR 7303), Marseille, France
Rôle dans le projet : Responsable du projet

Financeur(s) du projet et type de financement

2011 : Fonds incitatif recherche, Aix Marseille université : 10 000 euros. Le financement a permis l'élaboration du site accueillant la base de données, la reproduction numérique de lettres de/à Séguier conservées hors de Nîmes, l'organisation d'une journée d'études.

2012-2017 : utilisation des crédits IUF d'Emmanuelle Chapron (à la hauteur de 25 000 euros environ)

2020 : financement CAHIER (3000 euros)

Référence de la convention de financement

Aucune convention de financement

Institution / organisme / unité porteuses du projet

Laboratoire TELEMMe (Temps, Espaces, Langages, Europe méridionale-Méditerranée, UMR 7303), CNRS - Aix-Marseille Université

Partenaires (identifier les organismes partenaires, ressources et co-financeurs du projet)

Le projet d'édition et d'étude des archives savantes de Séguier est une initiative conjointe du laboratoire Telemme et de l'Institut européen Séguier (association nîmoise fondée en 2005).

Descriptif et objectif(s) du projet

Voir : https://cahier.hypotheses.org/seguier

Centré sur le savant nîmois Jean-François Séguier (1703-1784), le projet « Écritures savantes au siècle des Lumières » est une contribution à l'histoire des mobilités et de la communication intellectuelle dans l'Europe du XVIIIe siècle. Cet antiquaire et botaniste a en effet laissé une importante correspondance (environ 3500 lettres, dont un tiers provenant de savants non-régnicoles) et attiré dans son cabinet d'antiquités et d'histoire naturelle de nombreux voyageurs européens (près de 1500 pour la seule décennie 1770). Au-delà de l'éclairage que cette figure d'érudit méridional peut apporter sur le fonctionnement de la vie savante à l'écart des grandes capitales culturelles, une telle entreprise s'inscrit dans les réflexions actuelles sur ces deux modalités complémentaires de faire connaissance et travailler que sont, à l'époque moderne, la correspondance et le voyage.

Le projet consiste en l'édition numérique de la correspondance de Jean-François Séguier (1703-1784). Les relations épistolaires entretenues par l'antiquaire et botaniste nîmois avec de très nombreux savants européens, ainsi qu'avec tout un milieu d'amateurs méridionaux, ouvrent de nombreuses perspectives de recherche. Elles permettent d'interroger les formes du travail intellectuel à distance, la circulation des plantes, des graines et des livres, les cultures épistolaires (choix de la langue, civilités de l'entrée en correspondance), l'acculturation de différents milieux sociaux aux pratiques scientifiques (herborisation, collection, description des spécimens, manipulation des références bibliographiques…). Ces perspectives ont été illustrées par un colloque international (**Emmanuelle Chapron, François Pugnière (éd.)**, *Écriture épistolaire et production des savoirs au XVIIIe siècle. Les réseaux de Jean-François Séguier*, Paris, Classiques Garnier, 2019).

Pour mener à bien ce projet, un groupe de recherche, le Comité international Séguier (CIS), a été constitué et s'est réuni pour la première fois en mars 2010. Il était placé par convention sous la double tutelle de l'UMR 73030 Telemme (CNRS-Université de Provence, aujourd'hui Aix Marseille Université) et de l'Institut européen Séguier (association loi 1901, fondée en 2005), qui coordonnait depuis sa fondation les recherches autour du savant nîmois. Le CIS rassemblait des historiens, des historiens et philosophes des sciences, des littéraires de différents pays européens et était organisé en trois comités :

1) le **Comité scientifique**, sous la présidence de Brigitte Marin (Professeur d'histoire moderne, Université de Provence-MMSH), chargé du pilotage scientifique général du projet, est constitué de :

- Gabriel Audisio (Professeur émérite d'histoire moderne, Université de Provence / Institut européen Séguier)
- Jean Boutier (Directeur d'études à l'EHESS, Centre Norbert Elias, Marseille)
- Laurence Brockliss (Professeur d'histoire moderne, Université d'Oxford)
- Marina Caffiero (Professeur d'histoire moderne, Université La Sapienza, Rome)
- Michel Christol (Professeur d'histoire ancienne, Paris-I Sorbonne)
- Willem Frijhoff (Professeur d'histoire moderne, Université libre d'Amsterdam)

- Sergey Karp (Centre d'étude du XVIII[e] siècle, Académie des sciences, Russie)
- Hans-Jürgen Lüsebrink (Professeur d'histoire moderne, Université de Sarrebrück)
- Daniel Roche (Professeur honoraire, Collège de France)

2) le **Comité de recherche**, sous la présidence d'Emmanuelle Chapron (MCF en histoire moderne, Université de Provence-Telemme), chargé des recherches sur le terrain et de l'incrémentation du site, est constitué de :

- Arnaud Bartolomei (MCF en histoire moderne, Université de Nice)
- Robert Chamboredon (Professeur agrégé d'histoire-géographie, Nîmes / Institut Séguier)
- Samuel Cordier (docteur du Muséum national d'histoire naturelle, en poste à Genève)
- Ivano Dal Prete (Visiting Scholar, Lecturer, Yale University)
- Claire Davison-Pégon (Professeur de littérature, Université de Provence)
- Véronique Krings (MCF en histoire romaine, Université Toulouse-II)
- Gilles Montègre (MCF en histoire moderne, Université Grenoble-II)
- François Pugnière (Professeur d'histoire-géographie, Nîmes / IES)
- David Rousseau (doctorant en histoire moderne sous la direction de Pierre-Yves Beaurepaire, Université de Nice)
- Anne Saada (CNRS-UMR 8547 Pays germaniques : histoire culture philosophie).

3) le **Comité d'organisation**, sous la présidence de Gabriel Audisio (Institut européen Séguier, Professeur d'histoire émérite de l'Université de Provence) coordonne les activités des groupes et assure le relais avec l'IES. Il comprend :

- Evelyne Bret (Conservateur chargée des fonds anciens de la Bibliothèque municipale de Nîmes)
- Hélène Deronne (MCF en histoire de l'art, Université d'Avignon)
- Jean-Marie Guillon (Professeur d'histoire, Université de Provence- UMR Telemme)
- Jean-Louis Meunier (président de l'Institut européen Séguier)
- Jean-Michel Ott (trésorier du CIS, Institut européen Séguier)
- Rüdiger Stephan (chargé des relations internationales à l'Institut européen Séguier)
- Julie Thérond (chargée de communication, Institut européen Séguier)

Ce comité international Séguier n'a jamais réellement fonctionné. Il ne s'est jamais réuni en assemblée plénière après cette première rencontre et n'a été que rarement sollicité.

Après le délitement puis l'arrêt des activités de l'Institut européen Séguier (association loi 1905), l'entreprise s'est poursuivie sous la direction conjointe d'Emmanuelle Chapron (maître de conférences puis professeur d'histoire moderne, Aix Marseille Université), d'Eric Carroll (ingénieur de recherche en informatique, CNRS, Telemme) et de François Pugnière (professeur d'histoire-géographie à Nîmes, membre associé de l'EA 4424 CRISES, Montpellier III), avec le seul soutien de l'UMR Telemme.

Dates et durée

Date de début de financement et de début des travaux : 2010
Date de fin de financement et de fin des travaux : 2025 (fin des travaux estimée)

Mots clés du projet

- Correspondance savante
- Édition numérique (Édition électronique)

- Collections
- Archives

Publications (articles, pré-proposition, site web, …)

Site web du projet : www.seguier.org (le lien n'est plus actif)

En plus du site, pour une valorisation plus rapide du projet, un blog a été créé sur la plateforme Hypothèses en septembre 2015. Il permet de mettre en évidence les dernières publications sur le site, des points d'avancement, des réflexions théoriques. Administré par Emmanuelle Chapron, il compte une soixantaine de billets.

Carnet Hypothèse dédié au projet Séguier : https://seguier.hypotheses.org/

Listes des publications liées au projet :

- 2011 : journée d'études « Conserver, archiver, éditer. Usages de la correspondance savante, xvii^{e}-xviii^{e} siècles ». Actes publiés dans la Bibliothèque de l'École des chartes, 2013 [2017] sous la direction d'Emmanuelle Chapron et Jean Boutier.
- 2016 : colloque international Savoirs à l'œuvre, savants au travail. Actes publiés dans le volume **Emmanuelle Chapron et François Pugnière (dir.)**, *Écriture épistolaire et production des savoirs au xviii^{e} siècle. Les réseaux de Jean-François Séguier*, Paris, Classiques Garnier, 2019, 315 p.

Communications individuelles, Emmanuelle Chapron :

- Journée d'étude HISTARA, Paris, 2021 ;
- Colloque Condorcet, ENS, Paris, 2018 ;
- Séminaire d'Histoire des relations internationales, Poitiers, 2018 ;
- Assemblée du consortium CAHIER, Paris, 2016.

Publications individuelles :

- **Emmanuelle Chapron**, « Monde savant et ventes de bibliothèques en France méridionale dans la seconde moitié du xviii^{e} siècle », *Annales du Midi*, 283, 2013, p. 409-429 [halshs-01487318]. 2013
- **Emmanuelle Chapron**. *L'Europe à Nîmes : les carnets de Jean-François Séguier (1732-1784)*, Editions Barthelemy, 2008

3. Présentation et description du corpus

- ***Présentation de la section***

Cette section décrit le corpus et ses données. Elle décrit de façon plus précise les données du projet, les méthodes appliquées pour les collecter, etc. Ici, nous décrivons les données du Consortium CAHIER dans leur ensemble mais vous trouverez en annexe des exemples plus précis basés sur des projets membres de CAHIER

- ***Recommandations***

Il s'agira de préciser le mode de collecte et l'origine des données, les centres d'archives, bibliothèques ou centres d'études hébergeant les données y compris si les données procèdent d'un moissonnage de ressources en ligne. L'organisation du corpus, l'arborescence des fichiers, le système de nommage et de gestion des répertoires et des fichiers doit être décrite. De même que la nature des données, leurs formats, leur volumétrie (en poids et nombre de fichiers), leur état, etc. Pour que les données soient réutilisables sur le long terme, les formats doivent être ouverts et non propriétaires et les données stockées dans des entrepôts accessibles.

Nom du projet

Écritures savantes au siècle des Lumières. La correspondance et les carnets de visiteurs de Jean-François Séguier

Présenter et décrivez le corpus

Le cœur du corpus visé par le projet est la correspondance de Séguier, qu'on se propose d'éditer intégralement. Sa partie passive (correspondance reçue) est conservée pour l'essentiel à la Bibliothèque Carré d'Art de Nîmes et à la Bibliothèque nationale de France. Les lettres écrites par le Nîmois sont dispersées dans de nombreuses bibliothèques et fonds d'archives européens, notamment en Allemagne, en Suisse, au Royaume-Uni et en Italie. Leur recensement a supposé de longues recherches dans les catalogues et inventaires en ligne, et parfois des déplacements in situ. L'état du recensement des lettres et de leur transcription est tenu à jour sur le blog du projet : https://seguier.hypotheses.org/category/etats-des-lieux

Dans un second temps, il est prévu d'associer au projet la transcription du reste des archives de Séguier conservées à la Bibliothèque Carré d'Art de Nîmes (carnet de voyage, notes inédites, mémoires, catalogue de bibliothèque, inventaire des collections).

Environ 3710 lettres ont été repérées à ce jour :

La correspondance active représente 822 lettres et la correspondance passive, 2888 lettres.

Parmi cet ensemble, 2700 lettres sont conservées à la Bibliothèque du Carré d'art de Nîmes, le reste est dispersé dans plus d'une cinquantaine d'établissements situés dans neuf pays (Allemagne, Autriche, France, Italie, Pays-Bas, Royaume-Uni, Suède, Suisse, Russie).

Plus de 2200 lettres sont accessibles sur www.seguier.org.

Dans un second temps, le site accueillera l'édition des papiers de travail du savant : carnets de travail et de voyage, répertoires des visiteurs, notes de lecture, croquis, inventaires et catalogues autographes des collections.

Période couverte par le corpus, auteur(s) concerné(s)

Auteur : Séguier, Jean-François (1703-1784)
ISNI : 0000000108147857
ARK BnF : http://catalogue.bnf.fr/ark:/12148/cb10576633j
IdRef : 07713320X
Période du corpus : 1703 - 1784

Organisation du corpus

Les informations sur les données sont rassemblées et organisées dans un tableur Excel unique qui servira à leur dépôt dans Nakala.
Les colonnes rassemblent les métadonnées de la lettre : titre; expéditeur; date de la lettre; lieu et pays d'expédition; récepteur; lieu et pays de réception; langue; établissement de conservation de la lettre; cote; transcription intégrale; annotations; dessins et croquis; autres mentions descriptives; thèmes; personnes citées; ouvrages cités; contributeur [chercheur qui a transcrit la lettre]; éditeur scientifique; références bibliographiques; lien vers Gallica [si manuscrit numérisés]; droits sur l'image; licence.
Les lignes correspondent aux lettres. Chaque lettre est identifiée par un ID numérique progressif (de ID1 à ID 2263) qui permet de faire le lien avec les images de la lettre.
Les images sont nommées par cet ID et un numéro progressif (par exemple : ID263_1, ID263_2, etc.). Si on ne dispose pas des droits sur l'image, la lettre est liée à un fichier image générique "Image non disponible".

Mode de collecte et origine des données

Lettres détenues en majorité par la bibliothèque du Carré d'art de Nîmes. Elles ont été numérisées et transcrites afin d'être accessibles à la fois en mode image et en mode texte.

Etat du corpus numérique (types et natures des données, modifications effectuées sur les données, volumétrie)

Le corpus des lettres est estimé à 3500 lettres. Actuellement, 2200 ont été transcrites et environ 500 sont en cours de transcription ou de versement dans la base de données. Des recherches *in situ* devraient être entreprises pour retrouver et transcrire le reste des lettres, notamment en Russie et en Italie.

La convention liant l'Institut européen Séguier et l'UMR 7303 Telemme prévoyait la mise à disposition d'un ingénieur de recherche en informatique, Eric Carroll, pour la mise en place d'un site internet qui permette d'interroger et de consulter les données. Ce site a été conçu en deux temps :

1) Dans un premier temps (2010-2013), une interface de travail a été construite pour permettre une saisie collaborative des données (métadonnées des lettres, transcription intégrale et indexation des textes, chargement des images). Elle était fondée sur une base de données relationnelle, comprenant une vingtaine de tables, animée par une instance de SqlServer 2008 R2. La deuxième couche consistait en un développement d'une application internet riche (Ajax, JavaScript, dot Net 4.0, C#, XHTML/CSS, DublinCore/OAI, UML).

2) Dans un second temps (2013), une interface publique a été mise en place par une entreprise privée (Walter Wizman), en raison des multiples obligations professionnelles d'Eric Carroll au sein du laboratoire. A partir de l'adresse www.seguier.org, il était désormais possible d'interroger la lettre à partir de requêtes par nom (auteur de la lettre, destinataire, individus cités dans la lettre), par date et lieu d'expédition, par institution de conservation, par thème. Il était aussi possible d'effectuer des requêtes en texte intégral ou de sélectionner les lettres accompagnées de croquis ou auxquelles sont

jointes des objets, livres, graines et plantes ou petites antiquités. La liste des résultats permettait d'accéder aux éléments d'identification de la lettre et à son texte intégral, éventuellement à son image numérique, selon les conventions passées avec les établissements.

Après le départ d'Eric Carroll, qui a quitté l'UMR Telemme en février 2020, le site www.seguier.org n'a plus été entretenu, le nom de domaine n'a plus été payé et le site n'est plus accessible.

Les données textuelles ont été récupérées sous la forme d'un tableur Excell d'environ 2200 lignes et 30 colonnes. Avec les images des lettres (5000 fichiers image), l'ensemble a été stocké sur les serveurs de la MMSH et sur le Sharedocs du consortium CAHIER. Le consortium a mis à disposition du projet deux stagiaires (Andrés Echevarria Pelaes et Ala Eddine) chargés d'accompagner le dépôt des données sur Nakala.

Métadonnées, créées et standards et formats utilisés

Les métadonnées descriptives, administratives et techniques

Métadonnées de la lettre :

- Descriptives : titre ; expéditeur ; date de la lettre ; lieu et pays d'expédition ; récepteur ; lieu et pays de réception ; langue ; établissement de conservation de la lettre ; cote ; transcription intégrale ; annotations ; dessins et croquis ; autres mentions descriptives ; thèmes ; personnes citées ; ouvrages cités ; références bibliographiques
- Administratives : contributeur [chercheur qui a transcrit la lettre] ; éditeur scientifique
- Techniques : lien vers Gallica [si manuscrit numérisés] ; droits sur l'image ; licence.

Les métadonnées structurelles et l'annotation sémantique

Pas de données d'enrichissement sémantique ou de métadonnées structurelles

Référentiels d'indexation utilisés (vocabulaires contrôlés - thésaurus ou ontologies disciplinaires - et/ou indexation libre)

Indexation à partir d'une liste propre au projet.

4. Modalités de partage, de sauvegarde et de protection des données. Volumétrie des données stockées et espaces choisis.

- ***Présentation de la section***

Cette section décrit la documentation produite au cours projet. Il s'agit d'une documentation autre que numérique (sur support papier par exemple). Si elle existe, il est important de la décrire. Cette section décrit également les lieux et infrastructures de stockage des données pendant le projet.

- ***Recommandations***

Il s'agira de préciser ici le matériel physique et les lieux de stockage des données. Idéalement, il faudrait stocker les données dans au moins 2 endroits, éviter le stockage externe et privilégier les outils mis à disposition par l'institution. Pour cela, il peut être nécessaire de savoir quel est le volume approximatif des données à sauvegarder, l'espace de stockage nécessaire, la périodicité des sauvegardes, le nom et la nature du service fourni par l'institution, etc. On peut également indiquer les procédures de sauvegarde mises en place (fréquence des sauvegardes, automatisée ou non ?), les personnes en charge de la protection de ces données et du contrôle de l'accès, le mode de récupération des données en cas d'incident…

Accès, partage et limites des données

Les données sont stockées actuellement sur les serveurs de la MMSH et sur le Sharedocs du consortium CAHIER. Elles sont en cours de dépôt sur Nakala. Un site est en cours de construction sur Nakala-web, l'accès sera libre.

La publication des images a fait l'objet de conventions particulières avec les établissements de conservation, qui précisent les mentions légales à apporter. Les conventions devraient être renouvelées avant la publication des données sur Nakala. Les images des lettres conservées par la BnF (entre autres) ne sont pas couvertes par ces conventions.

Les transcriptions sont sous licence CC_BY 4.0

5. Responsabilités et ressources pour la gestion des données

- ***Présentation de la section***

Cette section décrit, identifie, présente et nomme les responsables de la gestion des données.

- ***Recommandations***

Afin de respecter les principes FAIR, CAHIER recommande le dépôt de celles-ci dans l'entrepôt Nakala (https://www.nakala.fr/). Ce service de dépôt et de stockage des données est proposé par la TGIR HumaNum pour les SHS. Il assure la gestion pérenne et sûre des données. Utiliser Nakala n'empêche pas de recourir à un second dépôt sur un autre entrepôt ou sur une plateforme institutionnelle.

Responsable de la gestion des données

CHAPRON, Emmanuelle, IdHAL : emmanuelle-chapron ; ORCID : https://orcid.org/0000-0001-9907-7961, Aix-Marseille Université, CNRS - TELEMMe (UMR 7303), Marseille, France
Rôle dans le projet : Responsable du projet

Évaluation des coûts (budgets, personnels et temps) dédiés à rendre les données FAIR (temps et budgets pour la collecte et la diffusion des données, pour le stockage et l'archivage).

Dans le cadre du Consortium CAHIER, les moyens assumés par l'infrastructure Huma-Num ont concerné les tâches suivantes :

- mise à disposition de moyens matériels tels que des serveurs, machines virtuelles, logiciels dédiés et licences supplémentaires dont les coûts et abonnements ne sont pas supportés par les projets, soit une économie estimée à ~5000€ / an pour chaque projet membre
- mise à disposition de moyens humains (ETP) pour des tâches spécifiques relevant à la fois de la gestion des moyens matériels (serveurs, machines, etc.), du stockage des données et des actions de formation, soit une économie estimée à plus de ~50000€ / an pour chaque projet membre

Dans le cadre du projet Séguier, le consortium a accordé un financement à hauteur de 3 000€. Un espace de stockage et de partage de fichiers sharedocs a également été ouvert (service HumaNum).

Mise à disposition de personnel :

- La convention liant l'Institut européen Séguier et l'UMR 7303 Telemme prévoyait la mise à disposition d'un ingénieur de recherche en informatique, Eric Carroll, pour la mise en place d'un site internet.
- Le consortium CAHIER a mis à disposition du projet deux stagiaires (Andrés Echevarria Pelaes et Ala Eddine) chargés d'accompagner le dépôt des données sur Nakala.

Le projet a rassemblé au fil des ans de nombreux contributeurs qui ont alimenté la base de données en corpus de lettres liés à leurs projets de master, de thèse ou de recherche :

- Lily Servel, étudiante du master Histoire d'Aix Marseille Université, a participé au projet pendant son master et par un CDD de 2 mois en 2016
- Adeline Danerol, étudiante du master Histoire d'Aix Marseille Université, a participé au projet pendant son master et a effectué plusieurs missions ponctuelles pour le projet
- Etienne Stockland, doctorant

- Meike Knittel, doctorante
- Florence Catherine, enseignante du secondaire
- Gilles Montègre, MCF Université Grenobles Alpes
- Andrea Bruschi, docteur en histoire, rémunéré (auto-entrepreneur) pour la saisie de lettres italiennes en 2014 et 2021
- Claire Torreilles
- Véronique Chapron, retraitée bénévole

6. Archivage des données

Présentation de la section

Cette section décrit les données à conserver à court, moyen et long terme, les éventuelles données à détruire ou à laisser sous embargo et indique la durée de cette restriction.

Recommandations

A l'issue du projet, des jeux de données se prêteront à une conservation à long terme pour une utilisation future, tandis que d'autres données ne nécessitent qu'une préservation à moyen terme car jugées moins essentielles et au potentiel de réutilisation limité, voire, elles pourront être destructibles pour des raisons de légalité ou de confidentialité.

Plateforme pour l'archivage pérenne des données

En cours de dépôt sur Huma-Num, la question de l'archivage n'est pas encore évoquée. Toutefois, grâce au dépôt sur Nakala, Huma-Num devrait gérer l'archivage pérenne.

7. Partage des données à l'issue du projet

- ***Présentation de la section***

Cette section décrit la politique de dissémination des données. Elle indique s'il existe des limites à la diffusion des données, comment les données pourront être trouvées et réutilisées par les pairs, voire par le grand public.

- ***Recommandations***

Une bonne dissémination des données requiert, dans la mesure du possible, le respect des principes FAIR : les données doivent être trouvables (findables), accessibles, interopérables et réutilisables. Pour être réutilisables, les données doivent être faciles d'accès, identifiables et citables grâce à des identifiants uniques (DOI) et leur usage facilité par l'accompagnement d'une description et de documentations, par des formats ouverts et non propriétaires et par une disponibilité facilitée par un lieu de stockage (entrepôt) ouvert, gratuit et référencé par les moteurs de recherche.

Publications sur les données destinées à en améliorer l'exposition

La valorisation et la documentation sur les données sont rassemblées sur le carnet Hypothèses : inventaires régulièrement mis à jour des lettres repérées et transcrites, éclairages sur certaines lettres ou ensemble de lettres, cartes, mise en relation avec d'autres archives, annonce des publications, etc. Le blog continuera à fonctionner après le dépôt sur Nakala mais nous espérons pouvoir rapatrier et diffuser une partie des informations sur le nouveau site.

Conditions de réutilisation : licences et contrats pour l'ensemble du projet

Voir item -4-

E-Stampages

1. Plan de gestion de données (PGD) du projet E-Stampages

- ***Présentation de la section***

Cette section décrit le PGD : elle présente l'auteur du PGD, les relecteurs du PGD, les autres intervenants assurant la gestion du PGD et, le cas échéant, ses mises à jour.

- ***Recommandations***

Il est utile de désigner un responsable du PGD qui sera la personne à contacter. Il n'est pas nécessairement le responsable scientifique du projet. Il est recommandé d'associer ce responsable à son identifiant ORCID, IdRef, ISNI, IdHal et de nommer l'ensemble des personnes ayant contribué à la rédaction et à la relecture du PGD.

Le PGD évolue au fur et à mesure de l'avancée du projet de recherche et de l'enrichissement des données. Afin de faciliter sa rédaction, il est conseillé d'en produire une première version au début du projet, qui sera modifiée éventuellement en cours de projet, ainsi qu'à la fin du projet et d'indiquer les versions du PGD dans leur ordre antéchronologique en commençant par l'actuelle.

Auteures du plan de gestion des données :

BRUNET, Michèle, IdHAL : michele-brunet ; ORCID : https://orcid.org/0000-0003-1818-5237, Université Lyon 2, Histoire et Sources des Mondes Antiques (HiSoMA, UMR 5189), Lyon, France
Rôle dans le projet : Responsable du projet

L'HERMITE, Laurène, IdRef : https://www.idref.fr/236176927 ; Université de La Rochelle, Centre de recherches en histoire internationale et atlantique (EA1163), La Rochelle, France
Rôle dans le projet : co-auteure du PGD

Version du plan de gestion des données :

PGD V1: 30/10/2021, PGD projet E-Stampages
2 versions de ce PGD sont actuellement prévues

2. Présentation du projet et responsabilités

- ***Présentation de la section***

Cette section décrit le projet ou le corpus sur lequel porte le PGD. Elle décrit le projet, ses objectifs, participants, etc. Ici, nous décrivons le Consortium CAHIER mais vous trouverez en annexe des exemples plus précis basés sur des projets membres de CAHIER

- ***Recommandations***

Si le nom du projet est un acronyme, indiquez également la version développée.

Exemple : Antonomaz ("ANalyse auTOmatique et NumérisatiOn des MAZarinades")

Identifier la/le responsable scientifique du projet : Nom, Prénom, Institution, Laboratoire, Unité de rattachement, Ville, Pays. Mettre en lien son identifiant ORCID *(ou ISNI, IdRef, IdHal, ...). Si possible indiquez des données de contact (courriel, téléphone professionnel)*

*Exemple : Karine ABIVEN (*https://orcid.org/0000-0001-9518-1040*), Sens-Texte-Informatique-Histoire (STIH), EA 4509, Université Paris - Sorbonne, Paris IV, France*

Précisez également si le projet s'inscrit dans une programmation scientifique financée et les axes scientifiques liés à cette programmation :
Axes scientifique d'un Labex
Programme de financement d'un projet ANR, H2020
Axe ou programme scientifique d'une structure de recherche liée au porteur ou à l'équipe projet...

Nom du projet

E-STAMPAGES, Archivage à long terme et création d'une Bibliothèque numérique d'estampages grecs.

Responsable du projet (principal researcher) et unité de rattachement

BRUNET, Michèle, IdHAL : michele-brunet ; ORCID : https://orcid.org/0000-0003-1818-5237, Université Lyon 2, Histoire et Sources des Mondes Antiques (HiSoMA, UMR 5189), Lyon, France

Financeur(s) du projet et type de financement

Projet financé dans le cadre de l'appel à projets Bibliothèque Scientifique Numérique 5 (BSN5)

Institution / organisme / unité porteuses du projet

École française d'Athènes, porteur et gestionnaire dans le cadre de l'appel BSN 5, Athènes, Grèce

Partenaires (identifier les organismes partenaires, ressources et co-financeurs du projet)

Le projet s'effectue au sein d'un consortium financé dans le cadre de l'appel à projets 2014 Bibliothèque Scientifique Numérique 5 (BSN5), pour une durée de 18 mois, janvier 2015-juin 2016.
Membres du consortium :

- École française d'Athènes, porteur et gestionnaire dans le cadre de l'appel BSN 5, Athènes, Grèce
- Laboratoire Histoire et Sources des Mondes Antiques, HiSoMA, UMR 5189 du CNRS, Lyon, France

- Pôle Système d'Information et Réseau de la Maison de l'Orient et de la Méditerranée - Jean Pouilloux, Lyon, France
- The Digital Epigraphy & Archaeology Project, Université de Floride, Gainsville, Floride, USA
- (Depuis 2017) Laboratorio di Epigrafia Greca du Dipartimento di Studi Umanistici de l'Università Ca' Foscari, Venise, Italie

Descriptif et objectif(s) du projet

D'après : https://cahier.hypotheses.org/e-stampages et https://www.e-stampages.eu/s/e-stampages/page/projet

Création d'une bibliothèque numérique d'environ 6000 estampages d'inscriptions grecques sélectionnés sur des critères communs, issues des collections de l'EFA et d'HiSoMA et en provenance des sites de Thasos, Délos, Delphes et Philippes. Les images sont disponibles en version 2D et 3D, l'outil de visualisation diverses fonctionnalités destinées à faciliter le travail des épigraphistes, en particulier un variateur d'ombrage.
Objectifs :

- un archivage à long terme des originaux sous une forme dématérialisée (images .tiff), à des fins de conservation
- la diffusion en libre accès sur le Web de cette documentation scientifique essentielle pour le travail des épigraphistes, associant aux vues 2D et 3D tout un ensemble de métadonnées modélisées et structurées. L'intention est donc de créer une ressource documentaire uniquement centrée sur les estampages : il ne s'agit ni de rééditer des textes épigraphiques ni de les commenter.

Les métadonnées documentaires récupérées sur les bases de données existantes ont été modélisées, enrichies et enregistrées dans des formats et langages standardisés pour le Web sémantique, afin d'en faciliter l'interopérabilité avec d'autres catégories de ressources, complémentaires pour l'étude des inscriptions grecques — éditions électroniques des textes en TEI/EpiDoc, photographies des inscriptions, systèmes d'information géographique, etc., qui pourront être ultérieurement reliées à l'ectypothèque E-STAMPAGES.

Dates et durée

Date de début de financement et de début des travaux : janvier 2015
Date de fin de financement et de fin des travaux : fin de financement en juin 2016. Toutefois le projet se poursuit, E-Stampages est inscrit dans l'axe prioritaire de recherche Outils numériques de la recherche du programme scientifique quinquennal de l'École française d'Athènes 2017-2021 et le site E-STAMPAGES est toujours en cours d'enrichissement et d'amélioration.

Mots clés du projet

- Estampage
- Epigraphie
- Bibliothèques numériques
- Métadonnées
- Visualisation
- Ectypothèque

Publications (articles, pré-proposition, site web, …)

Site web du projet : https://www.e-stampages.eu/

Listes des articles publiés par le projet :

Antonetti, Claudia, Michèle Brunet, Eloisa Paganoni. 'Collezioni Di Calchi Epigrafici: Una Nuova Risorsa Digitale'. *Axon no. 2* (23 December 2019). http://doi.org/10.30687/Axon/2532-6848/2019/02/004

Brunet, Michèle, "E-stampages : la mise en ligne des collections d'estampages. Une nouvelle ressource pour l'étude des inscriptions grecques", *Comptes rendus des séances de l'Académie des Inscriptions et Belles-Lettres vol. 2019*, 2021

Bozia, Eleni, "Assessing the role of digital libraries of squeezes in epigraphic studies", In *Digital and Traditional Epigraphy in Context. Proceedings of the EAGLE 2016 International Conference on Digital and Traditional Epigraphy in Context. 27-29 January 2016. Rome, Italy*. p373-378, Sapienza Università Editrice, 2016 https://doi.org/10.13133/978-88-9377-021-7

Bozia, Eleni, "Augmenting the workspace of epigraphists: an interaction design study", in *Digital and Traditional Epigraphy in Context. Proceedings of the EAGLE 2016 International Conference on Digital and Traditional Epigraphy in Context. 27-29 January 2016. Rome, Italy*. p171-182, Sapienza Università Editrice, 2016 https://doi.org/10.13133/978-88-9377-021-7

Adeline Levivier, Elina Leblanc et Michèle Brunet, « E-STAMPAGES : archivage et publication en ligne d'une ectypothèque d'inscriptions grecques », *Les nouvelles de l'archéologie* [En ligne], 145 | 2016. http://journals.openedition.org/nda/3801 DOI : https://doi.org/10.4000/nda.3801

Bozia, Eleni, Barmpoutis, Angelos, Wagman, Robert S., "OPEN-ACCESS EPIGRAPHY, Electronic Dissemination of 3D-digitized Archaeological Material", in *Information Technologies for Epigraphy and Cultural Heritage: Proceedings of the first EAGLE International Conference,* Paris, 2014 https://f-origin.hypotheses.org/wp-content/blogs.dir/31/files/2014/09/Open-Access-Epigraphy.pdf

3. Présentation et description du corpus

- ***Présentation de la section***

Cette section décrit le corpus et ses données. Elle décrit de façon plus précise les données du projet, les méthodes appliquées pour les collecter, etc. Ici, nous décrivons les données du Consortium CAHIER dans leur ensemble mais vous trouverez en annexe des exemples plus précis basés sur des projets membres de CAHIER

- ***Recommandations***

Il s'agira de préciser le mode de collecte et l'origine des données, les centres d'archives, bibliothèques ou centres d'études hébergeant les données y compris si les données procèdent d'un moissonnage de ressources en ligne. L'organisation du corpus, l'arborescence des fichiers, le système de nommage et de gestion des répertoires et des fichiers doit être décrite. De même que la nature des données, leurs formats, leur volumétrie (en poids et nombre de fichiers), leur état, etc. Pour que les données soient réutilisables sur le long terme, les formats doivent être ouverts et non propriétaires et les données stockées dans des entrepôts accessibles.

Nom du projet

E-STAMPAGES, Archivage à long terme et création d'une Bibliothèque numérique d'estampages grecs.

Présenter et décrivez le corpus

L'estampage est l'empreinte d'une inscription réalisée sous forme d'un moulage à l'aide d'un papier vergé sans colle. Ce matériel est indispensable aux épigraphistes pour se constituer une collection de sources consultable à distance, pratique à stocker et à transporter.

Le corpus est constitué d'environ 6000 estampages issus des collections de l'Ecole Française d'Athènes (EFA) et de l'HiSoMa, sélectionnés selon plusieurs critères pour être publiés sur https://www.e-stampages.eu/.

- **La collection de l'EFA**

Dans le cadre de la première phase du programme E-STAMPAGES, ont été retenus pour numérisation et diffusion sur le web :

- les estampages les plus anciens et les estampages endommagés, dont la préservation est prioritaire

- les estampages « historiques », provenant des sites archéologiques dont l'exploration a été confiée à l'EFA depuis le XIXe siècle

- les estampages reliés aux corpus d'inscriptions dont l'EFA est l'éditeur scientifique

Collection	Dates extrêmes	Volumétrie
Delphes	1896-1987	2662
Thasos	1907-2014	1101

Délos	1903-1980	894
Béotie	1884-1891	218
Philippes	1914-1984	130
Asie Mineure	1884-1886	99
Chalcidique	1891	8
Etolie	1885	2
Crète	1889	1

- **La collection d'HiSoMa**

Dans le cadre de la première phase du programme E-STAMPAGES, ont été retenus pour numérisation et diffusion sur le web

- les estampages reliés à la collection de l'EFA par une histoire institutionnelle commune

- les estampages du fonds Jean Pouilloux, complémentaires des ensembles de Thasos et de Delphes conservés à l'École française d'Athènes

Collection	Dates extrêmes	Volumétrie
Phocide-Delphes	1975-1993	549
Thasos	1946-1956	449
Délos	1903-1925	3

Période couverte par le corpus, auteur(s) concerné(s)

Dates des estampages : 1884 - 2014

Organisation du corpus

Mode de collecte et origine des données

Voir https://www.e-stampages.eu/s/e-stampages/page/projet

Près de 6000 estampages d'inscriptions, provenant des quatre grands sites explorés par l'EFA Délos, Delphes, Thasos et Philippes, ont été numérisés et des vues 3D ont été créées, suivant le protocole développé par le Digital Epigraphy and Archeology project.

Les métadonnées documentaires, récupérées dans les bases de données préexistantes, ont été re-documentarisées (modélisées, enrichies et structurées) dans des formats standardisés, afin d'en faciliter l'interopérabilité à venir avec d'autres catégories de ressources, complémentaires pour l'étude des inscriptions grecques : éditions électroniques des textes en TEI/EpiDoc, photographies des inscriptions, systèmes d'information géographique, etc., qui seront reliées à l'ectypothèque E-STAMPAGES.

Types de données

Images (estampages) numérisées / photographies des inscriptions, format jpeg et png
Jeux de données textuelles en pdf
Transcriptions XML-TEI

Volumétrie

Entre 6000 et 6200 épigraphies numérisées

Autres données créées ou collectées pour documenter et/ou enrichir les corpus constitués.

Un thesaurus multilingue est élaboré grâce à l'outil OpenTheso, développé et maintenu au sein du réseau FRANTIQ par l'équipe de la Plateforme Tête de Réseaux Documentaires de la Maison de l'Orient et de la Méditerranée-Jean Pouilloux.

Métadonnées, créées et standards et formats utilisés

Les métadonnées sont récupérées des bases de données existantes, enrichies et redocumentarisées pour les rendre interopérables.
Standard EAD
Schémas RDF, XML
Ontologie CIDOC CRM

4. Modalités de partage, de sauvegarde et de protection des données. Volumétrie des données stockées et espaces choisis.

- ***Présentation de la section***

Cette section décrit la documentation produite au cours projet. Il s'agit d'une documentation autre que numérique (sur support papier par exemple). Si elle existe, il est important de la décrire. Cette section décrit également les lieux et infrastructures de stockage des données pendant le projet.

- ***Recommandations***

Il s'agira de préciser ici le matériel physique et les lieux de stockage des données. Idéalement, il faudrait stocker les données dans au moins 2 endroits, éviter le stockage externe et privilégier les outils mis à disposition par l'institution. Pour cela, il peut être nécessaire de savoir quel est le volume approximatif des données à sauvegarder, l'espace de stockage nécessaire, la périodicité des sauvegardes, le nom et la nature du service fourni par l'institution, etc.On peut également indiquer les procédures de sauvegarde mises en place (fréquence des sauvegardes, automatisée ou non ?), les personnes en charge de la protection de ces données et du contrôle de l'accès, le mode de récupération des données en cas d'incident…

Accès, partage et limites des données

Droits (voir https://www.e-stampages.eu/s/e-stampages/page/credits_rights)
Mise à disposition du contenu original créé ou redocumentarisé pour l'ectypothèque E-STAMPAGES

- Les images 2D des estampages
- Les images 3D des estampages
- Les métadonnées attachées aux estampages, aux inscriptions et aux artefacts

sont publiées selon les termes de la Licence Creative Commons Attribution - Partage dans les Mêmes Conditions 4.0 International

-- Sauf indication spécifique, les photographies proviennent pour la plupart de la photothèque de l'École française d'Athènes. Toutes sont créditées à leurs auteurs respectifs. Pour toute demande de reproduction imprimée, se reporter à la page contact sur le site de l'EFA
-- Artefacts supports des inscriptions : tous droits reservés aux Musées dépositaires

- Musées archéologiques de Grèce sous la tutelle du Ministère de la Culture et du Sport de Grèce - Ελληνική Δημοκρατία, Υπουργείο Πολιτισμού & Αθλητισμού
- Musée du Louvre, Paris

Pour la consultation des artefacts originaux, prendre contact avec les musées dépositaires.

5. Responsabilités et ressources pour la gestion des données

- ***Présentation de la section***

Cette section décrit, identifie, présente et nomme les responsables de la gestion des données.

- ***Recommandations***

Afin de respecter les principes FAIR, CAHIER recommande le dépôt de celles-ci dans l'entrepôt Nakala (https://www.nakala.fr/). Ce service de dépôt et de stockage des données est proposé par la TGIR HumaNum pour les SHS. Il assure la gestion pérenne et sûre des données. Utiliser Nakala n'empêche pas de recourir à un second dépôt sur un autre entrepôt ou sur une plateforme institutionnelle.

Responsable de la gestion des données :

BRUNET, Michèle, IdHAL : michele-brunet ; ORCID : https://orcid.org/0000-0003-1818-5237, Université Lyon 2, Histoire et Sources des Mondes Antiques (HiSoMA, UMR 5189), Lyon, France

Évaluation des coûts (budgets, personnels et temps) dédiés à rendre les données FAIR (temps et budgets pour la collecte et la diffusion des données, pour le stockage et l'archivage).

Dans le cadre du Consortium CAHIER, les moyens assumés par l'infrastructure Huma-Num ont concerné les tâches suivantes:

- mise à disposition de moyens matériels tels que des serveurs, machines virtuelles, logiciels dédiés et licences supplémentaires dont les coûts et abonnements ne sont pas supportés par les projets, soit une économie estimée à ~5000€ / an pour chaque projet membre
- mise à disposition de moyens humains (ETP) pour des tâches spécifiques relevant à la fois de la gestion des moyens matériels (serveurs, machines, etc.), du stockage des données et des actions de formation, soit une économie estimée à plus de ~50000€ / an pour chaque projet membre

4 stages de Master pro ont été effectués en renfort de l'équipe projet :

- Elina Leblanc, Master Pro « Patrimoine écrit et Edition numérique », Université de Tours, stage de 6 mois (avril-septembre 2015). Réflexion sur la structuration des métadonnées et leur modélisation en relation avec les ontologies et les formats standards, choix du CMS Omeka pour la diffusion.
- Evita Dionysopoulou, Master Pro « Archives et Images », Université Jean Jaurès Toulouse, stage de 4 mois (avril-juillet 2016). Traitement d'une partie du fonds conservé à HiSoMA dit « Fonds Homolle », inventaire des 4857 estampages, identification et saisie des métadonnées de 1282 documents, numérisation de 650. Création d'une exposition virtuelle (V.0) sur Homolle et le travail de l'épigraphiste avec le CMS Omeka.
- Hélène Vuidel, Master Pro Sibist, Enssib, stage de 4 mois (février-mai 2017) : contribution à la création d'un thesaurus sous OpenTheso.

Rémy Ienco, Master Pro Métiers de la science des Patrimoines, CESR Université de Tours (stage 1 en convention avec CNRS/UMR 5189 Hisoma, 280 heures, mars-mai 2021 puis stage 2, 210 heures, juin-juillet 2021, en convention avec le Musée du Louvre) : préparation des métadonnées avant intégration dans le CMS pour la série Thasos, mise en œuvre du module de cartographie Mapping, finalisation du Thesaurus EpiVoc sous OpenTheso (commun aux programmes E-stampages et IGLouvre)

Ichtya

1. Plan de gestion de données (PGD) du projet Ichtya

- **Présentation de la section**

Cette section décrit le PGD : elle présente l'auteur du PGD, les relecteurs du PGD, les autres intervenants assurant la gestion du PGD et, le cas échéant, ses mises à jour.

- **Recommandations**

Il est utile de désigner un responsable du PGD qui sera la personne à contacter. Il n'est pas nécessairement le responsable scientifique du projet. Il est recommandé d'associer ce responsable à son identifiant ORCID, IdRef, ISNI, IdHal et de nommer l'ensemble des personnes ayant contribué à la rédaction et à la relecture du PGD.
Le PGD évolue au fur et à mesure de l'avancée du projet de recherche et de l'enrichissement des données. Afin de faciliter sa rédaction, il est conseillé d'en produire une première version au début du projet, qui sera modifiée éventuellement en cours de projet, ainsi qu'à la fin du projet et d'indiquer les versions du PGD dans leur ordre antéchronologique en commençant par l'actuelle.

Auteurs du plan de gestion des données :

Gauvin, Brigitte, ISNI : 0000000061551536, Université de Caen-Normandie, CRAHAM (UMR 6273), Caen, France
Rôle dans le projet : co-responsable du projet

Buquet, Thierry, IdHal : thierry-buquet, Orcid : 0000-0003-2956-8217 ; CNRS, CRAHAM (UMR 6273), Caen, France
Rôle dans le projet : co-responsable du projet

Buard, Pierre-Yves, Pôle Document numérique, MRSH, Université de Caen Normandie – CNRS (USR 3486), Caen, France
Rôle dans le projet : Responsable technique et éditorial

L'HERMITE, Laurène, IdRef : https://www.idref.fr/236176927 ; Université de La Rochelle, Centre de recherches en histoire internationale et atlantique (EA1163), La Rochelle, France
Rôle dans le projet : co-auteure du PGD

Version du plan de gestion des données :

PGD V1: 30/10/2021, PGD projet Ichtya
2 versions de ce PGD sont actuellement prévues.

2. Présentation du projet et responsabilités

- **Présentation de la section**

Cette section décrit le projet ou le corpus sur lequel porte le PGD. Elle décrit le projet, ses objectifs, participants, etc. Ici, nous décrivons le Consortium CAHIER mais vous trouverez en annexe des exemples plus précis basés sur des projets membres de CAHIER

- **Recommandations**

Si le nom du projet est un acronyme, indiquez également la version développée.
Exemple : Antonomaz ("ANalyse auTOmatique et NumérisatiOn des MAZarinades")
Identifier la/le responsable scientifique du projet : Nom, Prénom, Institution, Laboratoire, Unité de rattachement, Ville, Pays. Mettre en lien son identifiant ORCID (ou ISNI, IdRef, IdHal, ...). Si possible indiquez des données de contact (courriel, téléphone professionnel)
Exemple : Karine ABIVEN (https://orcid.org/0000-0001-9518-1040), Sens-Texte-Informatique-Histoire (STIH), EA 4509, Université Paris - Sorbonne, Paris IV, France
Précisez également si le projet s'inscrit dans une programmation scientifique financée et les axes scientifiques liés à cette programmation :

- *Axes scientifique d'un Labex*
- *Programme de financement d'un projet ANR, H2020*
- *Axe ou programme scientifique d'une structure de recherche liée au porteur ou à l'équipe projet...*

Nom du projet

ICHTYA Corpus de traités latins d'ichtyologie et histoire des savoirs sur la faune aquatique

Responsable du projet (principal researcher) et unité de rattachement

Gauvin, Brigitte, ISNI : 0000000061551536, Université de Caen-Normandie, CRAHAM (UMR 6273), Caen, France
Rôle dans le projet : co-responsable du projet

Buquet, Thierry, IdHal : thierry-buquet, Orcid : 0000-0003-2956-8217 ; CNRS, CRAHAM (UMR 6273), Caen, France
Rôle dans le projet : co-responsable du projet

Buard, Pierre-Yves, Pôle Document numérique, MRSH, Université de Caen Normandie – CNRS (USR 3486), Caen, France
Rôle dans le projet : Responsable technique et éditorial

Financeur(s) du projet et type de financement

CRAHAM (financement du laboratoire)

Institution / organisme / unité porteuses du projet

CRAHAM - Université Caen - CNRS

Partenaires (identifier les organismes partenaires, ressources et co-financeurs du projet)

- GDRI Zoomathia consacré à l'étude de la transmission des savoirs zoologiques, Antiquité-Moyen Âge (dir. A. Zucker, CEPAM, université de Nice).
- Sourcencyme (Sources des Encyclopédies Médiévales), piloté par Isabelle Draelants, à l'Institut de recherche et d'histoire des textes (Paris).
- Le projet ICHTYA bénéficie du soutien informatique et éditorial de la Maison de recherche en sciences humaines (Pôle document numérique, resp. Pierre-Yves Buard), et des Presses universitaires de Caen.
- Consortium CAHIER (Corpus d'Auteurs pour les Humanités, Informatisation, Édition, Recherche. Littérature, philosophie, histoire de l'art) (TGIR Huma-Num)

Descriptif et objectif(s) du projet

L'objectif du projet Ichtya est la mise en ligne d'un corpus de traités latins d'ichtyologie, permettant d'apprécier le contenu du savoir zoologique véhiculé pendant l'Antiquité et le Moyen Âge, avant la publication des grands traités d'ichtyologie du XVIe siècle. Il se fonde sur la base de la *Biblioteca Ichthyologica* de Pierre Artedi (collaborateur de Linné et fondateur de l'ichtyologie moderne). Un des objectifs du projet Ichtya serait de reconstituer la *Biblioteca Ichthyologica* en la documentant et de façon générale, proposer une étude de l'histoire de la bibliographie ichtyologique.

La Bibliothèque Ichtya s'accompagne d'un thesaurus des noms de poissons latins et vernaculaires figurant dans les textes latins et d'une bibliographie spécifique établie sur zotero.org.

L'édition critique des traités d'ichtyologie médiévaux est réalisée en XML-TEI, permettant des publications multi-modales, consultables en ligne et disponibles sous forme de livre papier.

Dans sa première phase, le projet Ichtya est dédié aux textes de la période médiévale. Il prévoit d'associer des éditions critiques ponctuelles (paru : *Hortus sanitatis*, lib. 4 *De piscibus* ; en cours Thomas de Cantimpré, *Liber de natura rerum*, 6 – 7 ; Albert le Grand, *De animalibus*, 24) et une bibliothèque documentaire, la Bibliothèque Ichtya (textes sources ou similaires : Pline, nat. 9 et 32 ; Ambroise, *Hexameron*, 5 ; Basile, *Hexameron*, 7-8, Isidore de Séville *Etymologiae* 12, 6 ; Vincent de Beauvais, *Speculum naturale*, 17) dont les outils d'exploration doivent pouvoir être mutualisés. Dans des phases ultérieures, il s'agira a) de replacer ce matériel dans la perspective de la Bibliotheca ichtyologica de Peter Artedi en tenant compte de la documentation dont il disposait (éditions consultées), b) d'envisager les traités de la Renaissance. Un dernier axe de recherche concerne l'étude des synonymies et polyonimies entre les noms de poissons dans les traités ichtyologiques, en analysant comment les auteurs, à partir du Moyen Âge, indiquent ces équivalences de désignation et leurs méthodes d'identification zoologique.

Dates et durée

Date de début de financement et de début des travaux : 2009
Date de fin de financement et de fin des travaux : Projet en cours

Mots clés du projet

- poissons
- ichtyologie
- histoire de la zoologie / Zoologie – Histoire
- philologie
- édition critique

- édition électronique
- XML-TEI / Text Encoding Initiative
- EAD / Description archivistique encodée
- encyclopédies médiévales / Encyclopédies – Moyen âge
- latin

Publications (articles, pré-proposition, site web, …)

Page web descriptive du projet : https://www.craham.cnrs.fr/ichtya/

Sites et outils en ligne :
2020 - Mise en ligne de la Bibliothèque Ichtya : https://ichtya.unicaen.fr/
2020 - Mise en ligne du Thesaurus Ichtya des noms latins de poissons et de créatures aquatiques figurant dans les textes latins d'ichtyologie antique et médiévale : https://ichtya.unicaen.fr/lab/thesaurus/
2015 - Mise en ligne de la Bibliographie collaborative d'Ichtya : https://www.zotero.org/groups/356871/ichtya/library
2013 - Edition multimodale de l'*Hortus sanitatis* (livre des poissons) : http://www.unicaen.fr/puc/sources/depiscibus/accueil

Listes des articles publiés par le projet :
Voir https://www.craham.cnrs.fr/ichtya/

3. Présentation et description du corpus

- **Présentation de la section**

Cette section décrit le corpus et ses données. Elle décrit de façon plus précise les données du projet, les méthodes appliquées pour les collecter, etc. Ici, nous décrivons les données du Consortium CAHIER dans leur ensemble mais vous trouverez en annexe des exemples plus précis basés sur des projets membres de CAHIER

- **Recommandations**

Il s'agira de préciser le mode de collecte et l'origine des données, les centres d'archives, bibliothèques ou centres d'études hébergeant les données y compris si les données procèdent d'un moissonnage de ressources en ligne. L'organisation du corpus, l'arborescence des fichiers, le système de nommage et de gestion des répertoires et des fichiers doit être décrite. De même que la nature des données, leurs formats, leur volumétrie (en poids et nombre de fichiers), leur état, etc. Pour que les données soient réutilisables sur le long terme, les formats doivent être ouverts et non propriétaires et les données stockées dans des entrepôts accessibles.

Nom du projet

Ichtya - Corpus de traités latins d'ichtyologie et histoire des savoirs sur la faune aquatique

Présenter et décrivez le corpus

Le corpus constitutif de la bibliothèque numérique Ichtya rassemble des textes antiques et médiévaux latins consacrés à l'ichtyologie qui furent publiés dans l'Antiquité, au Moyen Âge et à la Renaissance. Elle s'inspire de la Bibliotheca Ichtyologica de Peter Artedi (1705-1735) et a pour vocation de mettre en ligne et à disposition des lecteurs un corpus latin consacré au savoir ichtyologique.

Le corpus a été océrisé et entièrement encodé en XML-TEI. Il est actuellement composé de 21 textes (8 en ligne publique) et 5 traductions (2 en ligne publique). Ces textes, attribués ou anonymes, sont de nature hétérogène (textes sacrés, encyclopédies, dialogues, poèmes) et de longueur très variable (fragments, chapitres, livres entiers).

Les textes sont annotés (indexation des zoonymes et des citations de sources). L'indexation des zoonymes s'appuie sur la constitution d'un index ichtyologique encodé en XML TEI.

Période couverte par le corpus, auteur(s) concerné(s)

Antiquité - Moyen Âge, Renaissance. Auteurs multiples

Organisation du corpus

Les textes sont classés en fonction de leur période de publication (Antiquité, Moyen-Age, Époque moderne) et les traductions font l'objet d'une section à part.

Mode de collecte et origine des données

Les données textuelles ont fait l'objet d'une campagne de reconnaissance de caractères.

Etat du corpus numérique

Extraits et textes complets encodés, pas d'images numérisées.

Types de données

Données textuelles transcrites et enrichies en XML-TEI

Volumétrie

2000 pages web en janvier 2020

Modifications effectuées sur les données, versions, …

Le corpus a été entièrement encodé en XML-TEI

Autres données créées ou collectées pour documenter et/ou enrichir les corpus constitués.

Le thesaurus est construit en XML-TEI. Il est composé d'autant de fichiers XML (notices) que de formes latines (au nominatif) ou vernaculaires rencontrées dans le corpus de la bibliothèque Ichtya. Chaque notice présente toujours la référence précise à la source dans laquelle le terme apparaît. Cette indication de source s'accompagne, autant que possible, d'une ou plusieurs identifications et, pour les appellations latines et grecques, de la référence scientifique qui valide ces identifications. Ces identifications peuvent être accompagnées d'une note de commentaire. Les notices peuvent aussi présenter deux sortes de renvois sous forme de liens : d'une part à la forme principale en cas de paronymie, de variante orthographique ou de forme vernaculaire, indication qui figure en tête de la fiche, à la place de l'identification ; de l'autre aux autres termes désignant le même animal sous un autre nom. L'indexation par le biais du format XML permet de faire des liens directement d'une forme à l'autre. Ce thesaurus fournit un outil de première utilité pour l'étude des synonymies et polyonymies entre les noms de poissons dans les traités ichtyologiques.

Chaque forme de nom de poisson ou créature aquatique rencontré dans le corpus de la bibliothèque Ichtya fait donc l'objet d'une notice XML-TEI et chaque occurrence est liée à une notice du thesaurus.

Génération de graphes RDF : À partir de l'encodage en arbre XML-TEI, des graphes ont pu être générés pour chaque notice, permettant de mettre en évidence les liens établis par les chercheurs : identification ; variantes graphiques ; notices en relation.

Ce système d'enrichissement sémantique permet la création et la visualisation de réseaux de notices qui se révèlent un matériel intéressant d'exploitation du corpus pour les chercheurs.

Métadonnées, créées et standards et formats utilisés

Les métadonnées descriptives, administratives et techniques

Descriptions des fonds en format XML selon les recommandations de la TEI.

Les métadonnées structurelles et l'annotation sémantique

L'ensemble du corpus (textes et thesaurus) est encodé en XML TEI avec un schéma dédié. L'ensemble de l'outillage développé et sa documentation sont accessibles à cette adresse : https://www.unicaen.fr/recherche/mrsh/document_numerique/outils/compilations

Référentiels d'indexation utilisés (vocabulaires contrôlés - thésaurus ou ontologies disciplinaires - et/ou indexation libre)

Indexation avec l'appui d'un thésaurus spécifique créé au sein du projet.
Lien vers le thesaurus de zoologie ancienne Thezoo (http://web.cepam.cnrs.fr/opentheso/)

4. Modalités de partage, de sauvegarde et de protection des données. Volumétrie des données stockées et espaces choisis.

- **Présentation de la section**

Cette section décrit la documentation produite au cours projet. Il s'agit d'une documentation autre que numérique (sur support papier par exemple). Si elle existe, il est important de la décrire. Cette section décrit également les lieux et infrastructures de stockage des données pendant le projet.

- **Recommandations**

Il s'agira de préciser ici le matériel physique et les lieux de stockage des données. Idéalement, il faudrait stocker les données dans au moins 2 endroits, éviter le stockage externe et privilégier les outils mis à disposition par l'institution. Pour cela, il peut être nécessaire de savoir quel est le volume approximatif des données à sauvegarder, l'espace de stockage nécessaire, la périodicité des sauvegardes, le nom et la nature du service fourni par l'institution, etc.On peut également indiquer les procédures de sauvegarde mises en place (fréquence des sauvegardes, automatisée ou non ?), les personnes en charge de la protection de ces données et du contrôle de l'accès, le mode de récupération des données en cas d'incident…

Accès, partage et limites des données

La mise à disposition des fichiers XML est prévue et dans ce cas l'intégralité des données sera moissonnable et interopérable.
Le dépôt des données dans Nakala est prévu dans l'année à venir.

5. Responsabilités et ressources pour la gestion des données

- **Présentation de la section**

Cette section décrit, identifie, présente et nomme les responsables de la gestion des données.

- **Recommandations**

Afin de respecter les principes FAIR, CAHIER recommande le dépôt de celles-ci dans l'entrepôt Nakala (https://www.nakala.fr/). Ce service de dépôt et de stockage des données est proposé par la TGIR HumaNum pour les SHS. Il assure la gestion pérenne et sûre des données. Utiliser Nakala n'empêche pas de recourir à un second dépôt sur un autre entrepôt ou sur une plateforme institutionnelle.

Responsable de la gestion des données

Buard, Pierre-Yves, Pôle Document numérique, MRSH, Université de Caen Normandie – CNRS (USR 3486), Caen, France

Rôle dans le projet : Responsable technique et éditorial

Évaluation des coûts (budgets, personnels et temps) dédiés à rendre les données FAIR (temps et budgets pour la collecte et la diffusion des données, pour le stockage et l'archivage).

Dans le cadre du Consortium CAHIER, les moyens assumés par l'infrastructure Huma-Num ont concerné les tâches suivantes :

- mise à disposition de moyens matériels tels que des serveurs, machines virtuelles, logiciels dédiés et licences supplémentaires dont les coûts et abonnements ne sont pas supportés par les projets, soit une économie estimée à ~5000€ / an pour chaque projet membre
- mise à disposition de moyens humains (ETP) pour des tâches spécifiques relevant à la fois de la gestion des moyens matériels (serveurs, machines, etc.), du stockage des données et des actions de formation, soit une économie estimée à plus de ~50000€ / an pour chaque projet membre

6. Archivage des données

- **Présentation de la section**

Cette section décrit les données à conserver à court, moyen et long terme, les éventuelles données à détruire ou à laisser sous embargo et indique la durée de cette restriction.

- **Recommandations**

A l'issue du projet, des jeux de données se prêteront à une conservation à long terme pour une utilisation future, tandis que d'autres données ne nécessitent qu'une préservation à moyen terme car jugées moins essentielles et au potentiel de réutilisation limité, voire, elles pourront être destructibles pour des raisons de légalité ou de confidentialité.

Grâce au dépôt dans Nakala, Huma-Num devrait gérer l'archivage pérenne.

Lafabrev (La Fabrique de la Révolution)

1. Plan de gestion de données (PGD) du projet LAFABREV (La Fabrique de la Révolution)

- **Présentation de la section**

Rédaction du PGD démarrée par Cécile Andrisi-Brémon le 11/10/2021
Cette section décrit le PGD : elle présente l'auteur du PGD, les relecteurs du PGD, les autres intervenants assurant la gestion du PGD et, le cas échéant, ses mises à jour.

- **Recommandations**

Il est utile de désigner un responsable du PGD qui sera la personne à contacter. Il n'est pas nécessairement le responsable scientifique du projet. Il est recommandé d'associer ce responsable à son identifiant ORCID, IdRef, ISNI, IdHal et de nommer l'ensemble des personnes ayant contribué à la rédaction et à la relecture du PGD.
Le PGD évolue au fur et à mesure de l'avancée du projet de recherche et de l'enrichissement des données. Afin de faciliter sa rédaction, il est conseillé d'en produire une première version au début du projet, qui sera modifiée éventuellement en cours de projet, ainsi qu'à la fin du projet et d'indiquer les versions du PGD dans leur ordre antéchronologique en commençant par l'actuelle.

Auteures du plan de gestion des données

ANDRISI-BRÉMON, Cécile, IdHAL : 1084562 ; IdRef : 193185725 ; ORCID : https://orcid.org/0000-0002-5130-339X, Autoentrepreneuse pour le compte de l'Université de Paris, en lien avec le CERILAC (UPR n°441), Paris, France ; ingénieure d'études en CDD d'octobre 2015 à juin 2019 (Centre Seebacher, CERILAC, Université Paris Diderot)

Rôle dans le projet : coordinatrice numérique (mises en ligne et relectures, en collaboration avec Paule Petitier et Olivier Ritz). Rédaction et compilation à partir de documents rédigés par Paule Petitier

PETITIER, Paule, IdHAL : 720408 ; ISNI : 0000000117567255 ; IdRef : 033175756 , Université de Paris, CERILAC (UPR 441), Paris, France

Rôle dans le projet : porteuse du projet et responsable scientifique

RITZ, Olivier, IdHAL : olivier-ritz ; ORCID : 0000-0001-5492-9403, Université de Paris, CERILAC (UPR 441), Paris, France

Rôle dans le projet : chercheur et coordinateur scientifique et numérique

L'HERMITE, Laurène, IdRef : https://www.idref.fr/236176927 ; Université de La Rochelle, Centre de recherches en histoire internationale et atlantique (EA1163), La Rochelle, France

Rôle dans le projet : co-auteure du PGD

Version du plan de gestion des données

PGD V1 : 19/10/2021, PGD projet LAFABREV
PGD V2 : 26/11/2021, PGD projet LAFABREV
Trois versions de ce PGD sont actuellement prévues.

2. Présentation du projet et responsabilités

- **Présentation de la section**

Cette section décrit le projet ou le corpus sur lequel porte le PGD. Elle décrit le projet, ses objectifs, participants, etc. Ici, nous décrivons le Consortium CAHIER mais vous trouverez en annexe des exemples plus précis basés sur des projets membres de CAHIER

- **Recommandations**

Si le nom du projet est un acronyme, indiquez également la version développée.
Exemple : Antonomaz ("ANalyse auTOmatique et NumérisatiOn des MAZarinades")
Identifier la/le responsable scientifique du projet : Nom, Prénom, Institution, Laboratoire, Unité de rattachement, Ville, Pays. Mettre en lien son identifiant ORCID (ou ISNI, IdRef, IdHal, ...). Si possible indiquez des données de contact (courriel, téléphone professionnel)
Exemple : Karine ABIVEN (https://orcid.org/0000-0001-9518-1040), Sens-Texte-Informatique-Histoire (STIH), EA 4509, Université Paris - Sorbonne, Paris IV, France
Précisez également si le projet s'inscrit dans une programmation scientifique financée et les axes scientifiques liés à cette programmation :

- *Axes scientifique d'un Labex*
- *Programme de financement d'un projet ANR, H2020*
- *Axe ou programme scientifique d'une structure de recherche liée au porteur ou à l'équipe projet...*

Nom du projet

La Fabrique de la Révolution (LAFABREV)

Responsable du projet (principal researcher) et unité de rattachement

PETITIER, Paule, IdHAL : 720408, Université de Paris, CERILAC (UPR 441), Paris, France
Rôle dans le projet : porteuse du projet et responsable scientifique

Autre coordinateur scientifique et numérique du projet : :

RITZ, Olivier, IdHAL : olivier-ritz ; ORCID : 0000-0001-5492-9403, Université de Paris, CERILAC (UPR 441), Paris, France
Rôle dans le projet : chercheur et coordinateur scientifique et numérique

Membres du projet :

Voir la galerie de membres consultable dans la rubrique « Équipe » du site, où figurent une partie des membres : https://lafabrev-michelet.lac.univ-paris-diderot.fr/equipe

ANDRISI-BRÉMON, Cécile (voir « Auteur du plan de gestion des données » et fiche à la rubrique « Équipe »)

ARNAUD, Émilien (Masterant en histoire, Master de recherche Humanités numériques et Computationnelles, École nationale des Chartes - PSL - ENS - EPHE - EHESS, détenteur d'un précédent Master de recherche en histoire de l'Université Paris 1 Panthéon-Sorbonne - Institut d'Histoire de la Révolution française - Institut d'Histoire Moderne et Contemporaine : travail sur l'index des références bibliographiques)

BERKERY, Charlotte (Docteure en littérature française du XIXe siècle : transcriptions)

FAURE, Claudie (Transcriptrice bénévole, ancienne chargée de recherche Laboratoire Traitement et Communication de l'Information LTCI-CNRS : nombreuses transcriptions. Voir fiche à la rubrique « Équipe »)

LALLIER, Thomas (Développeur web indépendant : développement du site LAFABREV. Voir fiche à la rubrique « Équipe »)

LEBARBÉ, Thomas (Conseiller scientifique et numérique : professeur en Humanités numériques à l'Université Grenoble Alpes et coordinateur du Consortium CAHIER auquel est rattaché le projet LAFABREV)

MAGRET, Maryelle (Relectrice-correctrice indépendante : préparation des fichiers XML et complétion des métadonnées à partir des volumes conservés à la BHVP, avant transcription des papiers par les transcripteurs)

MASEDA, Oriane (Détentrice d'un Master 1 Humanités numériques de l'École des chartes : nombreuses transcriptions)

MILOT-PINSON, Cécile (Cursus d'historienne ; désormais professeure des écoles : coordination numérique du projet jusqu'à début 2016 ; mise en place, avec son frère Baptiste Milot, développeur web, de l'outil numérique «chaîne éditoriale» élaboré par Thomas Lebarbé pour le projet «Manuscrits de Stendhal» ; formation des transcripteurs et rédaction d'un manuel d'encodage synthétisé par la suite dans un protocole de transcription récapitulatif)

MASSIP, Luc (Détenteur d'un Master 2 Lexicographie, Terminographie et Traitement Automatique de Corpus LTTAC de l'Université de Lille : correction automatisée d'erreurs dans les fichiers XML en Python et développements sur le site internet (allègement des pages ; mise en place d'une navigation entre entrées d'index et papiers ; mise en place d'une recherche par expressions régulières de type regex ; ajout de la page « Téléchargements » et de la rubrique « Équipe » . Voir fiche à la rubrique « Équipe »)

OUDAI CELSO, Yamina (PhD de l'Université de Venise : transcriptions. Voir fiche à la rubrique « Équipe »)

PETITIER, Paule (voir « Responsable du projet » et bientôt voir fiche à la rubrique « Équipe »)

RITZ, Olivier (voir « Autre coordinateur scientifique et numérique du projet » et voir fiche à la rubrique « Équipe »)

SAFA, Isabelle (Chercheuse, agrégée de lettres modernes et docteur en littérature française, autrice d'une thèse sur le roman historique d'A. Dumas, enseignante en lycée et classes préparatoires à Lille : transcriptions. Voir fiche à la rubrique « Équipe »)

WULF, Judith (Professeur à l'Université de Nantes, autrice d'une thèse sur V. Hugo : transcriptions)

Financeur(s) du projet et type de financement

- DIM STCN (Domaine d'intérêt majeur « Sciences du texte et connaissances nouvelles » de la région Île-de-France : crédits de fonctionnement pour payer les prestataires intervenant sur le projet (coordinatrice numérique et historien en charge de corriger l'index des références bibliographiques)

- UDPN (Usages des patrimoines numérisés) : crédits de personnel pour payer l'ingénieure d'études du projet (coordinatrice numérique) et des vacations de transcription (enseignants-chercheurs et doctorantes)
- Consortium CAHIER : crédits pour payer la numérisation des volumes effectuée par la société Digiscrib ; formations XLST pour l'ingénieure d'études du projet
- CERILAC, URP 441 (Centre d'Études et de Recherches Interdisciplinaires de l'UFR Lettres, Arts, Cinéma) de l'Université de Paris (ex-Paris Diderot) : crédits de personnel pour payer des vacations de transcription
- Université de Paris (ex-Paris Diderot) : crédits de personnel pour payer l'ingénieur d'études du projet (coordinatrice technique)

Référence de la convention de financement

n° DIMSTCN-2019-18

Institution / organisme / unité porteuses du projet

Université de Paris (ex-Paris Diderot)

Partenaires (identifier les organismes partenaires, ressources et co-financeurs du projet)

DIM STCN et Consortium CAHIER

Descriptif et objectif(s) du projet

Le programme de recherche LAFABREV vise à constituer un corpus numérique à partir des notes préparatoires de l'*Histoire de la Révolution française* de Jules Michelet.

Dates et durée

Date de début de financement et de début des travaux : 2014 /2015
Date de fin de financement et de fin des travaux : 2021/2022

Mots clés du projet

Jules Michelet ; Histoire ; Révolution française ; XVIIIe siècle ; XIXe siècle

Publications (articles, pré-proposition, site web, …)

Site web du projet : https://lafabrev-michelet.lac.univ-paris-diderot.fr/

L'outil de visualisation, présent sur le site internet du projet, a été développé par Thomas Lallier et amélioré par Luc Massip. Cet outil de visualisation propose un affichage du texte respectueux de ses principales caractéristiques structurelles et formelles, mais facilitant la lecture. Il comporte 7 index : index des personnes, des lieux, des États, des institutions, des références bibliographiques, des références littéraires et artistiques et des événements. Ces index comportent des hyperliens renvoyant vers des exemplaires numérisés des sources auxquelles Michelet fait référence dans ses papiers.

L'utilisateur peut naviguer entre les papiers et les index dans un sens et dans l'autre et effectuer des recherches avancées grâce à des expressions régulières de type regex.

Liste des articles publiés par le projet :

Articles de Paule Petitier :

- Grâce aux ressources numériques, on sait mieux comment travaillait Michelet, 21 janvier 2020.
- La Semaine sainte de Michelet. L'émergence de l'idée des fédérations à travers les papiers préparatoires de l'*Histoire de la Révolution française*, n° 46, 2018.

Article d'Olivier Ritz :

- « Renouveler la tradition : sources et références dans l'*Histoire de la Révolution française* de Michelet » (à paraître).

Billets du carnet de recherche « Littérature et Révolution » d'Olivier Ritz

- Vers un catalogue numérique de la Révolution, 7 novembre 2018.
- « La faim passe du peuple au Roi ! », 19 juin 2018.
- Dans les petits papiers de Michelet, 21 juin 2017.
- « La Fabrique de la Révolution », 27 avril 2016.

Autres livrables (guides, recommandations, etc.) :

- Guide d'expressions régulières appliqué au projet LAFABREV, rédigé et intégré au site internet par Luc Massip (stage d'Humanités numériques) : https://lafabrev-michelet.lac.univ-paris-diderot.fr/regex
- Protocole de transcription intitulé « Tableau récapitulatif des éléments », rédigé par Cécile Milot-Pinson et actualisé par Cécile Andrisi-Brémon : bientôt téléchargeable à la page « Téléchargements » du site : https://lafabrev-michelet.lac.univ-paris-diderot.fr/telechargements

3. Présentation et description du corpus

- **Présentation de la section**

Cette section décrit le corpus et ses données. Elle décrit de façon plus précise les données du projet, les méthodes appliquées pour les collecter, etc. Ici, nous décrivons les données du Consortium CAHIER dans leur ensemble mais vous trouverez en annexe des exemples plus précis basés sur des projets membres de CAHIER

- **Recommandations**

Il s'agira de préciser le mode de collecte et l'origine des données, les centres d'archives, bibliothèques ou centres d'études hébergeant les données y compris si les données procèdent d'un moissonnage de ressources en ligne. L'organisation du corpus, l'arborescence des fichiers, le système de nommage et de gestion des répertoires et des fichiers doit être décrite. De même que la nature des données, leurs formats, leur volumétrie (en poids et nombre de fichiers), leur état, etc. Pour que les données soient réutilisables sur le long terme, les formats doivent être ouverts et non propriétaires et les données stockées dans des entrepôts accessibles.

Nom du projet

La Fabrique de la Révolution (LAFABREV)

Présenter et décrivez le corpus

Ce projet d'humanités numériques se concentre sur la transcription d'un corpus manuscrit qui n'a pas encore été dépouillé : les sept volumes de notes préparatoires à l'*Histoire de la Révolution française*, conservés à la Bibliothèque historique de la Ville de Paris.

La transcription et l'indexation de quelque 1800 de ces feuillets (sur environ 2000) est susceptible de mieux faire connaître les sources documentaires sur lesquelles l'historien s'est appuyé, mais surtout sa méthode de travail et le(s) type(s) d'usage(s) qu'il fait de ces sources.

Période couverte par le corpus, auteur(s) concerné(s)

Auteur : Jules Michelet ;

Dates d'écriture : de mars-avril 1846 à décembre 1868.

Organisation du corpus

Organisation en 7 volumes de notes (6 volumes de notes préparatoires à l'*Histoire de la Révolution française* + 1 volume spécial sur les sections parisiennes)

Mode de collecte et origine des données

Les notes préparatoires à l'*Histoire de la Révolution française* de Michelet n'avaient jamais été exploitées avant leur numérisation.

Les six volumes de notes conservés à la Bibliothèque historique de la Ville de Paris sous le nom « Histoire de la Révolution française », correspondent à un ensemble complexe de documents préparatoires, en partie réorganisés après coup en vue d'éventuelles réutilisations. D'un point de vue matériel, leur conservation préventive au sein de la Bibliothèque par leur montage sur onglet en recueils factices, a durablement perturbé la lisibilité et l'architecture de ces fragments, organisés au départ sous forme de liasses et de chemises hiérarchisées selon un classement organique.

Il s'est donc agi en premier lieu de restituer le plus fidèlement que possible l'architecture d'origine de ces milliers de fragments et d'annotations dont le format et plus encore le statut sont fort variables.

Les liens et associations entre certains fragments dessinent en effet de véritables dossiers préparatoires, qui permettent de reconstituer les séquences narratives ou démonstratives de l'*HRF.*

Enfin, il existe un septième volume de notes, isolées de longue date, sans doute par Michelet lui-même après 1871. Le recueil factice ainsi constitué contient la seule trace d'une documentation, provenant pour l'essentiel des sections parisiennes et témoignant donc du pouvoir « sans-culotte ». En effet, à la suite de l'incendie de l'Hôtel de Ville en mai 1871, les archives de la municipalité parisienne durant la Révolution ont irrémédiablement disparu. La transcription et la mise en ligne intégrale de ce manuscrit permettent d'associer à la figure et au travail de l'écrivain celui de l'archiviste, dédoublant ainsi les formes de transmission et de fabrication des mémoires : à travers ses propres notes de travail, Michelet n'est pas seulement un historien qui écrit la Révolution française. Il est aussi le médiateur qui contribue à l'archiver.

État du corpus numérique

Le corpus numérique est composé des images des notes ou « papiers » manuscrits scannés, des fichiers XML des transcriptions et de sept index : index des personnes, des lieux, des États, des institutions, des références bibliographiques, des références littéraires et artistiques et des événements.

Tous les papiers (sauf rares exceptions) ont été transcrits, encodés, corrigés au moins une fois (pour la majeure partie des papiers), et sont en cours de nouvelle relecture à partir d'un tableau de repérage d'erreurs dans les formes normalisées (élaboré par Luc Massip et en partie complété par Olivier Ritz et Cécile Andrisi-Brémon).

Types de données

Les métadonnées et les transcriptions sont regroupées à l'intérieur d'un même fichier XML : les métadonnées sont décrites en EAD et les transcriptions sont encodées selon un schéma spécifique au projet (DTD), qui s'inspire de celui du projet des « Manuscrits de Stendhal », copiloté par Thomas Lebarbé, et qui peut être au moins partiellement transposable en TEI.

Les images des notes (une image pour une note) affichées sur le site sont au format JPEG dans une résolution réduite et une archive contenant la totalité des images sera également bientôt téléchargeable en haute résolution (JPEG, 300 DPI), depuis la page « Téléchargements » du site : https://lafabrev-michelet.lac.univ-paris-diderot.fr/telechargements

Chaque image présente sur le site correspond à un « papier ». Les volumes de la BHVP rassemblent des papiers de différentes longueurs : certains très longs et d'autres plus courts, voire très courts, collés ensemble sur de grandes pages. Les papiers courts ont donc été découpés dans un logiciel de traitement de l'image afin de figurer sur le site. Tous les rectos et les versos non vierges ont été numérisés. Certains versos ne semblant pas présenter d'intérêt scientifique ont été écartés de la transcription.

Toutes les images des papiers figurent sur le site « Chaîne éditoriale » du projet, de même que les fichiers XML des transcriptions, la DTD et la CSS : https://ce-michelet.app.univ-paris-diderot.fr/

Tous les scans des folios tels qu'initialement numérisés (rassemblant donc parfois plusieurs papiers) sont conservés au format TIFF (résolution 400 DPI, mode RVB) sur un serveur dédié, hébergé par l'Université de Paris.

Les index du site sont générés depuis un tableur collaboratif, exporté au format .CSV puis .JSON. Les index obtenus sont consultables sur le site par un système d'onglets accessibles à partir du pied de page et s'ouvrent à partir de l'index des noms de personnes :
https://lafabrev-michelet.lac.univ-paris-diderot.fr/index-personnes
Ces index comportent des hyperliens renvoyant vers des exemplaires numérisés des sources auxquelles Michelet fait référence dans ses papiers. Un important travail de correction et de normalisation de l'index des références bibliographiques a été effectué par Émilien Arnaud.

Grâce au travail de Thomas Lallier, l'utilisateur du site peut, par un système d'infobulles, naviguer des papiers vers les entrées d'index. Grâce au travail de Luc Massip, l'utilisateur peut en outre naviguer des entrées d'index vers les entités nommées correspondantes encodées à l'intérieur des transcriptions et des entités nommées encodées vers l'entrée d'index correspondante. Ce système se base sur des interrogations de la base de données réalisées à l'aide d'expressions régulières de type regex.

Volumétrie

1873 papiers (1873 images et autant de transcriptions au format XML) et 7 index consultables sur le site

Modifications effectuées sur les données, versions, …

Les modifications apportées dans les fichiers XML sont encodées à l'intérieur des métadonnées de la manière suivante :
<item_transcription date_transcription="AAAA/MM/JJ"
modifications="transcription_initiale" nom_transcripteur="Prénom + Nom"/>
Les modifications peuvent être de 3 types : "metadonnees", "transcription initiale" et "corrections".
Un bloc de validation permet en outre de préciser la date et l'auteur de la correction finale et la date et le nom de la validatrice (Paule Petitier).

Les modifications de la DTD sont inscrites en début de fichier : version 1, version 2, version 3.1 à 3.7

Les modifications du protocole de transcription (tableau récapitulatif des éléments) suivent les modifications de la DTD.

Autres données créées ou collectées pour documenter et/ou enrichir les corpus constitués.

7 index

Métadonnées, créées et standards et formats utilisés

Métadonnées décrites en EAD principalement (sauf bloc de validation)

Les métadonnées descriptives, administratives et techniques

Balises de métadonnées utilisées :
<description_corpus>
<titre_corpus>
<titre_principal_corpus>
<sous_titre_corpus>

```
        <proprietaire>
 <programme_recherche>
        <nom_programme>
        <financeur>
        <logiciel>
  <caracteristiques_document>
        <cote_fichier_numerique>
        <cote_volume>
        <cote_folio>
        <cote_folio_A>
        <cote_folio_C>
        <cote_folio_BIC>
        <description_materielle>
        <dimensions>
                <largeur>
                <hauteur>
        <aspect>
        <verso type_verso="" />
        <scripteur forme_normalisee="">
        <transcription>
        <item_transcription date_transcription="AAAA/MM/JJ" modifications="metadonnees"
nom_transcripteur="Maryelle Magret" />
        <item_transcription date_transcription="AAAA/MM/JJ"
modifications="transcription_initiale" nom_transcripteur="Prénom + Nom" />
        <item_transcription date_transcription="AAAA/MM/JJ" modifications="corrections"
nom_transcripteur="Prénom + Nom" />
         <validation>
        <fichier_corrige date_finalisation_correction="AAAA/MM/JJ" nom_correcteur="Prénom +
Nom" />
        <a_valider valeur="OUI" />
        <fichier_valide date_validation="AAAA/MM/JJ" nom_validateur="Prénom + Nom" />
        <a_publier valeur="OUI" />
        </validation>
```

Les métadonnées structurelles et l’annotation sémantique

Voir protocole de transcription bientôt téléchargeable à la page « Téléchargements » du site

Référentiels d’indexation utilisés (vocabulaires contrôlés - thésaurus ou ontologies disciplinaires - et/ou indexation libre)

Indexation s’inspirant librement de la TEI et indexation complémentaire libre

4. Modalités de partage, de sauvegarde et de protection des données. Volumétrie des données stockées et espaces choisis.

- **Présentation de la section**

Cette section décrit la documentation produite au cours projet. Il s'agit d'une documentation autre que numérique (sur support papier par exemple). Si elle existe, il est important de la décrire. Cette section décrit également les lieux et infrastructures de stockage des données pendant le projet.

- **Recommandations**

Il s'agira de préciser ici le matériel physique et les lieux de stockage des données. Idéalement, il faudrait stocker les données dans au moins 2 endroits, éviter le stockage externe et privilégier les outils mis à disposition par l'institution. Pour cela, il peut être nécessaire de savoir quel est le volume approximatif des données à sauvegarder, l'espace de stockage nécessaire, la périodicité des sauvegardes, le nom et la nature du service fourni par l'institution, etc. On peut également indiquer les procédures de sauvegarde mises en place (fréquence des sauvegardes, automatisée ou non ?), les personnes en charge de la protection de ces données et du contrôle de l'accès, le mode de récupération des données en cas d'incident…

Répertoire "htdocs" en cours de téléchargement pour vérifier volume total du site LAFABREV

Accès, partage et limites des données

Partage des données selon la licence Creative Commons Attribution - Pas d'utilisation commerciale - Partage dans les mêmes conditions 4.0 International.

5. Responsabilités et ressources pour la gestion des données

- **Présentation de la section**

Cette section décrit, identifie, présente et nomme les responsables de la gestion des données.

- **Recommandations**

Afin de respecter les principes FAIR, CAHIER recommande le dépôt de celles-ci dans l'entrepôt Nakala (https://www.nakala.fr/). Ce service de dépôt et de stockage des données est proposé par la TGIR HumaNum pour les SHS. Il assure la gestion pérenne et sûre des données. Utiliser Nakala n'empêche pas de recourir à un second dépôt sur un autre entrepôt ou sur une plateforme institutionnelle.

Le projet LAFABREV envisage un stockage de ses données par la TGIR Huma-Num.

Évaluation des coûts (budgets, personnels et temps) dédiés à rendre les données FAIR (temps et budgets pour la collecte et la diffusion des données, pour le stockage et l'archivage).

Dans le cadre du Consortium CAHIER, les moyens assumés par l'infrastructure Huma-Num ont concerné les tâches suivantes :

- mise à disposition de moyens matériels tels que des serveurs, machines virtuelles, logiciels dédiés et licences supplémentaires dont les coûts et abonnements ne sont pas supportés par les projets, soit une économie estimée à ~5000€ / an pour chaque projet membre
- mise à disposition de moyens humains (ETP) pour des tâches spécifiques relevant à la fois de la gestion des moyens matériels (serveurs, machines, etc.), du stockage des données et des actions de formation, soit une économie estimée à plus de ~50000€ / an pour chaque projet membre

6. Archivage des données

- **Présentation de la section**

Cette section décrit les données à conserver à court, moyen et long terme, les éventuelles données à détruire ou à laisser sous embargo et indique la durée de cette restriction.

- **Recommandations**

A l'issue du projet, des jeux de données se prêteront à une conservation à long terme pour une utilisation future, tandis que d'autres données ne nécessitent qu'une préservation à moyen terme car jugées moins essentielles et au potentiel de réutilisation limité, voire, elles pourront être destructibles pour des raisons de légalité ou de confidentialité.

Plateforme pour l'archivage pérenne des données

Actuellement le site internet LAFABREV et la base de données qui lui est associée sont hébergés sur les serveurs de l'Université de Paris (ex-Paris Diderot).
La chaîne éditoriale et sa base de données ainsi que le serveur contenant les images scannées des folios des volumes de la BHVP sont également hébergés sur les serveurs de l'Université de Paris.

Si un dépôt sur Nakala était fait, la TGIR Huma-Num devrait gérer l'archivage pérenne.

Durée de conservation des données

Aussi longtemps que possible

Volume des données à conserver

Calcul du volume de données à conserver en cours

Coûts alloués à la conservation

Coûts à déterminer avec HumaNum si cette solution est choisie.

Outils, méthodes, procédures nécessaires pour accéder à ces données archivées et les réutiliser

Accès VPN au serveur (via client SFTP) et à la base de données (via phpMyAdmin)

7. Partage des données à l'issue du projet

- **Présentation de la section**

Cette section décrit la politique de dissémination des données. Elle indique s'il existe des limites à la diffusion des données, comment les données pourront être trouvées et réutilisées par les pairs, voire par le grand public.

- **Recommandations**

*Une bonne dissémination des données requiert, dans la mesure du possible, le respect des principes FAIR : les données doivent être trouvables (*findable*), accessibles, interopérables et réutilisables. Pour être réutilisables, les données doivent être faciles d'accès, identifiables et citables grâce à des identifiants uniques (DOI) et leur usage facilité par l'accompagnement d'une description et de documentations, par des formats ouverts et non propriétaires et par une disponibilité facilitée par un lieu de stockage (entrepôt) ouvert, gratuit et référencé par les moteurs de recherche.*

Potentiel de réutilisation des données

Fichiers XML transformables (au moins partiellement) en XML-TEI
Demande de DOI envisageable

Éléments d'accompagnement qui permettent la réutilisation des données.

Aide pour interpréter les données : protocole de transcription

Publications sur les données destinées à en améliorer l'exposition

Publications dans revues et carnet de recherche détaillées dans point "Publications" de la partie 2 de ce document

Conditions de réutilisation : licences et contrats pour l'ensemble du projet

Licence Creative Commons Attribution - Pas d'utilisation commerciale - Partage dans les mêmes conditions 4.0 International

Les dossiers de Bouvard et Pécuchet

1. Plan de gestion de données (PGD) du projet Les dossiers de Bouvard et Pécuchet

- **Présentation de la section**

Cette section décrit le PGD : elle présente l'auteur du PGD, les relecteurs du PGD, les autres intervenants assurant la gestion du PGD et, le cas échéant, ses mises à jour.

- **Recommandations**

Il est utile de désigner un responsable du PGD qui sera la personne à contacter. Il n'est pas nécessairement le responsable scientifique du projet. Il est recommandé d'associer ce responsable à son identifiant ORCID, IdRef, ISNI, IdHal et de nommer l'ensemble des personnes ayant contribué à la rédaction et à la relecture du PGD.
Le PGD évolue au fur et à mesure de l'avancée du projet de recherche et de l'enrichissement des données. Afin de faciliter sa rédaction, il est conseillé d'en produire une première version au début du projet, qui sera modifiée éventuellement en cours de projet, ainsi qu'à la fin du projet et d'indiquer les versions du PGD dans leur ordre antéchronologique en commençant par l'actuelle.

Auteures du plan de gestion des données

Dord-Crouslé, Stéphanie, IdHAL : stephanie-dord-crousle ; ORCID : 0000-0002-6683-9509, CNRS (UMR 5317 IHRIM), France
Rôle dans le projet : responsable scientifique

L'HERMITE, Laurène, IdRef : https://www.idref.fr/236176927 ; Université de La Rochelle, Centre de recherches en histoire internationale et atlantique (EA1163), La Rochelle, France
Rôle dans le projet : co-auteure du PGD

Version du plan de gestion des données :

PGD V1 : 30/10/2021
1 version de ce PGD est actuellement prévue

2. Présentation du projet et responsabilités

- **Présentation de la section**

Cette section décrit le projet ou le corpus sur lequel porte le PGD. Elle décrit le projet, ses objectifs, participants, etc. Ici, nous décrivons le Consortium CAHIER mais vous trouverez en annexe des exemples plus précis basés sur des projets membres de CAHIER

- **Recommandations**

Si le nom du projet est un acronyme, indiquez également la version développée.
Exemple : Antonomaz ("ANalyse auTOmatique et NumérisatiOn des MAZarinades")
Identifier la/le responsable scientifique du projet : Nom, Prénom, Institution, Laboratoire, Unité de rattachement, Ville, Pays. Mettre en lien son identifiant ORCID (ou ISNI, IdRef, IdHal, ...). Si possible indiquez des données de contact (courriel, téléphone professionnel)
Exemple : Karine ABIVEN (https://orcid.org/0000-0001-9518-1040), Sens-Texte-Informatique-Histoire (STIH), EA 4509, Université Paris - Sorbonne, Paris IV, France
Précisez également si le projet s'inscrit dans une programmation scientifique financée et les axes scientifiques liés à cette programmation :

- *Axes scientifique d'un Labex*
- *Programme de financement d'un projet ANR, H2020*
- *Axe ou programme scientifique d'une structure de recherche liée au porteur ou à l'équipe projet...*

Nom du projet

Les dossiers de Bouvard et Pécuchet

Responsable du projet (principal researcher) et unité de rattachement

Dord-Crouslé, Stéphanie, IdHAL : stephanie-dord-crousle ; ORCID : 0000-0002-6683-9509, CNRS (UMR 5317 IHRIM), France

Rôle dans le projet : responsable scientifique

Financeur(s) du projet et type de financement

Le projet a bénéficié d'un soutien financier spécifique

- du CNRS (appel d'offres « ATIP Jeunes chercheurs » 2006 du Département Sciences humaines et sociales)
- de l'ANR (appel à projets « Corpus et outils de la recherche en Sciences humaines et sociales » du programme Sciences humaines et sociales 2007 ; https://anr.fr/Projet-ANR-07-CORP-0009).
- de la Région Rhône-Alpes (allocation doctorale allouée au projet dans le cadre du Cluster de recherche n° 13 « Culture, patrimoine, création » 2007-2010)
- du Ministère des Affaires étrangères et européennes (Partenariat Hubert Curien Galilée 2009)
- de la TGIR Huma-Num par l'intermédiaire du consortium CAHIER (2012, 2013 et 2018).

Voir http://www.dossiers-flaubert.fr/projet-partenaires-soutiens

Institution / organisme / unité porteuses du projet

L'UMR 5317 IHRIM qui a pris la suite de l'UMR 5611 LIRE le 1er janvier 2016.

Partenaires (identifier les organismes partenaires, ressources et co-financeurs du projet)

Le projet – dans sa version initiale – a été réalisé entre 2006 et 2012

- dans le cadre de l'UMR 5611 LIRE (« Littérature, Idéologies, REprésentations, XVIIIe et XIXe siècles »), unité mixte de recherche qui associait le CNRS, l'Université Lumière - Lyon 2, l'Université Jean Monnet de Saint-Étienne, l'Université Stendhal - Grenoble 3 et l'ENS de Lyon
- avec le soutien technique de l'USR 3385 ISH (unité mixte de services de l'Institut des Sciences de l'Homme)
- de l'ENS-LSH (École normale supérieure - Lettres et sciences humaines) devenue l'ENS de Lyon (École normale supérieure de Lyon)
- et du TGE Adonis.

Il se poursuit aujourd'hui :

- dans le cadre de l'UMR 5317 IHRIM qui a pris la suite de l'UMR 5611 LIRE le 1er janvier 2016
- avec le soutien technique de la TGIR Huma-Num
- au sein du consortium CAHIER.

Voir http://www.dossiers-flaubert.fr/projet-partenaires-soutiens

Descriptif et objectif(s) du projet

Conservés à la bibliothèque municipale de Rouen, les dossiers de *Bouvard et Pécuchet*, le dernier roman – posthume et inachevé – de Gustave Flaubert (1821-1880), constituent un ensemble patrimonial imposant (2 400 feuillets), cohérent, d'importance scientifique et culturelle reconnue. Ils sont porteurs d'une dimension épistémologique singulière : composés pour rédiger une « encyclopédie critique en farce », ils proposent une configuration critique des savoirs au XIXe siècle, originale et révélatrice. Ils forment le socle de la présente édition. Mais d'autres dossiers existent ailleurs qui ont vocation à enrichir le site en rejoignant progressivement et virtuellement leurs semblables. Car c'est l'ensemble de ce chantier documentaire qui a servi à rédiger le premier volume de l'œuvre et aurait dû être réutilisé pour la composition d'un second volume, jamais écrit en raison de la mort soudaine du romancier.

Or, en raison de leur volume, de leur organisation complexe et indéfiniment mouvante, ainsi que de leurs contenus scientifiques extrêmement variés, les dossiers ne peuvent pas être édités de manière satisfaisante sous une forme imprimée. C'est particulièrement vrai pour les pages préparées en vue du second volume du roman : les annotations que l'écrivain y a portées, indiquant le lieu probable du classement, sont souvent plurielles et obligent à conserver aux fragments textuels une mobilité qui est nécessairement défaite par la fixité d'une édition imprimée.

Dépassant cette limite en recourant au support électronique et à l'encodage XML-TEI intégral du corpus, la présente édition offre l'accès :

- aux images, à la transcription (formats diplomatique et textuel) et aux métadonnées des pages du corpus,
- à un moteur de recherche plein texte,
- à trois bibliothèques permettant d'identifier les références utilisées par Flaubert et de circuler dans le corpus
- et à un outil de production de « seconds volumes » possibles : l'agenceur.

Date et durée

Date de début des travaux : 2006
Date de fin des travaux : non prévue (tant qu'il y aura des dossiers à intégrer)

Mots clés du projet

- Roman -- Dossiers documentaires ;
- Text Encoding Initiative (langage de balisage) ;
- Bibliothèques numériques ;
- Flaubert, Gustave (1821-1880) Bouvard et Pécuchet ;
- Œuvre inachevée ;
- Reconstitution conjecturale ;
- Edition posthume ;
- Manuscrits inédits ;

Publications (articles, pré-proposition, site web, …)

Site web du projet : http://www.dossiers-flaubert.fr
Le site Les dossiers de Bouvard et Pécuchet s'est vu attribuer un ISSN (International Standard Serial Number) par la Bibliothèque nationale de France : ISSN 2495-9979.
Sont ainsi soulignés et valorisés les enrichissements progressifs du site qui le constituent en ressource intégratrice.
Ressortent en particulier d'une publication en série :

- les reconstitutions conjecturales du « second volume » existant déjà sous forme papier ainsi que les agencements créés par des internautes (après validation par le comité scientifique du projet) qui seront progressivement mis en ligne sur le site
- et l'ajout à venir, sur la plateforme éditoriale, de nouveaux dossiers de notes non conservés à la bibliothèque municipale de Rouen.

Carnet de recherche du projet : https://flaubert.hypotheses.org/

Listes des articles publiés par le projet : https://halshs.archives-ouvertes.fr/ANR-07-CORP-009

Autres livrables (guides, recommandations, etc.) :

- **S. Dord-Crouslé**. Compte-rendu de fin de projet -Projet ANR-07-CORP-009 BOUVARD - *Les Dossiers de Bouvard et Pécuchet de Flaubert. Enrichissement, valorisation, documentation d'un corpus multi supports : Programme " Corpus et outils de la Recherche en Sciences Humaines et Sociales "* 2007. [Rapport de recherche] ANR (Agence Nationale de la Recherche - France). 2012. ⟨halshs-00760914⟩

3. Présentation et description du corpus

- **Présentation de la section**

Cette section décrit le corpus et ses données. Elle décrit de façon plus précise les données du projet, les méthodes appliquées pour les collecter, etc. Ici, nous décrivons les données du Consortium CAHIER dans leur ensemble mais vous trouverez en annexe des exemples plus précis basés sur des projets membres de CAHIER

- **Recommandations**

Il s'agira de préciser le mode de collecte et l'origine des données, les centres d'archives, bibliothèques ou centres d'études hébergeant les données y compris si les données procèdent d'un moissonnage de ressources en ligne. L'organisation du corpus, l'arborescence des fichiers, le système de nommage et de gestion des répertoires et des fichiers doit être décrite. De même que la nature des données, leurs formats, leur volumétrie (en poids et nombre de fichiers), leur état, etc. Pour que les données soient réutilisables sur le long terme, les formats doivent être ouverts et non propriétaires et les données stockées dans des entrepôts accessibles.

Présentez et décrivez le corpus

Voir http://www.dossiers-flaubert.fr/edition-corpus
Au corpus originel intégralement conservé à la bibliothèque municipale de Rouen s'ajoutent maintenant, peu à peu, les dossiers conservés ailleurs. En dépassant les strictes limites qu'il s'était d'abord fixées (l'édition d'un ensemble patrimonial cohérent conservé à la bibliothèque municipale de Rouen), le site commence à réaliser pleinement le projet scientifique d'ampleur qui est le sien : donner à lire l'ensemble de la documentation réunie par Flaubert pour son entreprise encyclopédique « en farce » et permettre à l'agenceur d'y puiser des matériaux – pour certains inconnus – en vue de la création ou de l'enrichissement de « seconds volumes » possibles.
Ont été ajoutés les dossiers Rousseau, Hegel et Mirabeau.

Période couverte par le corpus, auteur(s) concerné(s)

Auteur : Flaubert, Gustave (1821-1880)
ISNI : 0000000122762442
ARK BnF : http://catalogue.bnf.fr/ark:/12148/cb11902894q
IDRef : 026866552
Période du corpus : Les documents produits par Flaubert l'ont été entre 1860 et 1880.
Mais les ouvrages pris en note ont un empan chronologique bien plus large: antiquité - 1880.

Organisation du corpus

Unités constitutives du corpus
(voir http://www.dossiers-flaubert.fr/edition-unites-constitutives) :

- **La page** (unité matérielle) : L'unité constitutive matérielle du corpus des dossiers documentaires de Bouvard et Pécuchet est *la page*. Un feuillet est formé de deux pages ou folios, un recto et un verso – dont l'un peut être vierge.
- **Les transcriptions** (associées aux pages). Elles sont de 4 types :
  - La *transcription ultra-diplomatique* se présente sous la forme d'un fichier PDF généré à partir d'un logiciel de traitement de texte. Elle reprend toutes les particularités de la graphie du scripteur.
  - La *transcription diplomatique* (format HTML et générée à partir des fichiers XML/TEI) conserve tous les traitements textuels décrits pour la version ultra-diplomatique. En

revanche, elle homogénéise et rationalise une partie des dimensions topographiques et graphiques.
    - La *transcription normalisée* (format HTML et générée à partir des fichiers XML/TEI) achève d'homogénéiser le rendu topographique des pages en déterminant et en ne conservant que quelques espaces signifiants (essentiellement deux : la marge et le corps du texte). Mais surtout, elle propose un texte intelligible par tous les lecteurs, débarrassé des particularités et des graphies déviantes propres à chaque scripteur.
    - La *transcription enrichie* (format HTML et générée à partir des fichiers XML/TEI) permet de faire le lien entre les versions diplomatique et normalisée.
- **Les textes** (unités logiques à fondement matériel) : les pages regroupées selon un ordre validé scientifiquement forment des *textes* qui appartiennent à des catégories typologiques homogènes. Il s'agit d'un autre point d'entrée vers la lecture et l'exploitation des Dossiers. Techniquement, chaque texte, que ce soit en version diplomatique ou en version normalisée, présente l'agrégation – au sein d'une page HTML – du contenu balisé en XML/TEI de l'ensemble des pages concernées ; il est doté d'une URL spécifique et est accessible sur le site à partir d'une page de sommaire permettant de lister, type par type, la totalité des textes du corpus selon différents ordres (classement patrimonial, ordre alphabétique des titres, etc.)
- **Les fragments** : les pages sont composées de *fragments textuels*. Ce sont les unités logiques fondamentales de l'édition électronique du corpus : à leur niveau va être vérifiée et promue la mobilité des éléments constitutifs des dossiers documentaires de Bouvard et Pécuchet. La possibilité de créer des reconstitutions conjecturales du second volume du roman repose sur le découpage de l'intégralité du corpus en fragments textuels, opération qui le rend manipulable et infiniment réagençable. Chaque fragment textuel est accessible par l'intermédiaire d'une métadonnée (« Référence bibliographique de fragment ») attachée à la page où il apparaît, et élucidant la référence bibliographique exacte du fragment copié par Flaubert ou l'un de ses collaborateurs.
- **La citation** est le regroupement de tous les fragments présentant la réalisation textuelle de la même référence bibliographique. Chaque citation possède une page dédiée, pourvue d'une URL et présentant toutes les informations nécessaires à son identification.

Plan de nommage des fichiers :

Collection (nom)	**Collection (description)**	**Volume (nom)**	**Volume (description)**	
BnF	NAF	28825	"Littérature - esthétique"	
Montmorency	Musée JJ Rousseau	495	"Notes sur rousseau"	Montmorency_495_f_001_r.jpg
Rouen	BM	g 225-3	Feuillets épars	
[Rouen]		Volume 1 cote g 226-1 => à corriger en :	g 226-1	
	Information requise			

	uniquement dans le TEIHeader			
Antibes	Vente Caroline Franklin Groult, 1931	066	“Esthétique de Hegel”	Antibes_066_f_001_r

Mode de collecte et origine des données

Origine des images, manuscrits et autres pièces des dossiers :

- Bibliothèque municipale de Rouen

La numérisation du microfilm de sauvegarde des documents visés par le projet (microfilm acquis au prix public) nous a permis de constituer une base de 3500 images en noir et blanc de qualité médiocre .
Parallèlement, achat de près de 300 images HD couleur (sur 3500) au prix public grâce à une partie du financement reçu de l’ANR.
Manuscrit “définitif” Premier volume, notes, brouillons, plans, scénarios, notes de lecture.
Pages préparatoires Second volume
Puis mise à disposition à titre gracieux par la bibliothèque de la numérisation des deux dossiers concernant le *Dictionnaire des idées reçues* (soit 130 images).

- Musée Jean-Jacques Rousseau et Bibliothèque d’études rousseauistes, Montmorency : mise à disposition des images à titre gracieux par convention

- Antibes, vente Caroline Franklin Groult, 1931: images uniquement, issues de collections privées, non référencées.

Etat du corpus numérique, types de données et volumétrie

Le corpus est ouvert et en cours d’enrichissement. De nouveaux dossiers sont régulièrement ajoutés. L’encodage est toujours en cours et en phase d’amélioration.
Il contient :

- Base de données SQL : Données de travail du projet, références bibliographiques (actuellement plus de 20000), etc. Voir par exemple http://www.dossiers-flaubert.fr/index.php?node=bibliotheques – 100 Mo
- Base de données XML : Transcriptions TEI – 3500 transcriptions à terme ; 2000 disponibles actuellement
- Images : Fac-similés de manuscrits – 3500 images (toutes disponibles en ligne), 5 Go
- Images : Fragments d’images découpés (partiellement disponibles – pour 300 images) – 21000 images (prévision), 3Go

Modifications effectuées sur les données, versions, …

Transcription issue d’un traitement de texte puis balisage en TEI.

Autres données créées ou collectées pour documenter et/ou enrichir les corpus constitués.

Trois bibliothèques de références bibliographiques destinées à enrichir et lier les données.
Voir : http://www.dossiers-flaubert.fr/index.php?node=bibliotheques

L'agenceur :
Il s'agit d'un outil informatique de production de "seconds volumes" possibles. Pour utiliser cet outil, il faut préalablement s'identifier (se connecter ou, lors de sa première visite, créer un compte).
Cette démarche permet à chaque demandeur de disposer d'un espace de travail personnel et privé.
Des informations utiles à la prise en main de l'agenceur sont disponibles :

- sur la page "Agencements" de présentation de l'édition
- en cliquant sur le point d'interrogation ("?") qui se trouve en haut et à droite sur certaines pages de l'espace de travail
- ou bien en consultant les pages dédiées du carnet de recherche du projet qui comportent plusieurs tutoriels (par exemple, ici).

Vous pouvez aussi regarder la vidéo expliquant le fonctionnement du site et plus particulièrement celui de l'agenceur.

Métadonnées, créées et standards et formats utilisés

Les métadonnées sont entièrement accessibles sur le site des "Dossiers…".
Exemple : http://www.dossiers-flaubert.fr/cote-Antibes_066_f_002__r-meta
Les métadonnées ne sont pas standardisées et les champs d'indexation sont libres.
Des transcriptions XML/TEI et des descriptions sont toujours en cours de réalisation.

Les métadonnées descriptives, administratives et techniques

Cote, Scripteur du manuscrit, ensemble textuel d'où provient l'extrait, nom du transcripteur.

Les métadonnées structurelles et l'annotation sémantique

Chronologie du document, provenance, classement typologique* et caractéristiques matérielles.
**Les métadonnées de classement* : pour chaque page sont proposés un classement typologique (en fonction des différents types de pages qui existent dans le corpus : notes de lecture, pages préparées pour le second volume, documentation brute imprimée, etc.) ; un classement chronologique (selon la datation plus ou moins précise qui peut être affectée à chaque page en fonction d'informations internes, comme les filiations génétiques, ou externes, la date d'emprunt d'un ouvrage consignée dans le registre d'une bibliothèque ou la mention, dans une lettre, de la période à laquelle une lecture a été faite par le romancier) ; et un classement par scripteur (Flaubert est évidemment le plus largement représenté, mais bien d'autres personnes lui ont apporté leur aide et ont laissé des traces manuscrites dans les dossiers de Rouen, au premier rang desquelles son ami Edmond Laporte, mais aussi son « disciple » Guy de Maupassant). Ces classements permettent de proposer trois points d'accès au corpus qui s'ajoutent à celui que fournit, par défaut, le classement patrimonial, accessible par les sections du descriptif établi par l'institution de conservation ou par cotes.

Annotations ou métadonnées d'enrichissement

Annotations critiques.
Transcriptions TEI destinées à enrichir les manuscrits : transcription ultra diplomatique, diplomatique, transcription normalisée, transcription enrichie.
Références bibliographiques

4. Modalités de partage, de sauvegarde et de protection des données. Volumétrie des données stockées et espaces choisis.

- **Présentation de la section**

Cette section décrit la documentation produite au cours projet. Il s'agit d'une documentation autre que numérique (sur support papier par exemple). Si elle existe, il est important de la décrire. Cette section décrit également les lieux et infrastructures de stockage des données pendant le projet.

- **Recommandations**

Il s'agira de préciser ici le matériel physique et les lieux de stockage des données. Idéalement, il faudrait stocker les données dans au moins 2 endroits, éviter le stockage externe et privilégier les outils mis à disposition par l'institution. Pour cela, il peut être nécessaire de savoir quel est le volume approximatif des données à sauvegarder, l'espace de stockage nécessaire, la périodicité des sauvegardes, le nom et la nature du service fourni par l'institution, etc. On peut également indiquer les procédures de sauvegarde mises en place (fréquence des sauvegardes, automatisée ou non ?), les personnes en charge de la protection de ces données et du contrôle de l'accès, le mode de récupération des données en cas d'incident…

Stockage

Actuellement les données sont stockées via les services dédiés d'Huma-Num (Huma-Num Box ?)
Le transfert des données sur l'entrepôt Nakala (service Huma-Num) est prévu sous peu.

Volume des données stockées (qui sera également celui des données à sauvegarder) :

- PHP/JavaScript/CSS...: 892 Mo
- MySQL: 125 Mo
- SOLR: 38 Mo
- XML/TEI: 83 Mo
- Images (des manuscrits sous différents formats): 4,9 Go

Accès, partage et limites (d'accessibilité) des données

Une collection de 900 pages est moissonnée par Isidore

Concernant certaines images et manuscrits issus de collections de bibliothèques ou de fonds privés, il est à prévoir des règles de partage et de réutilisation :

Pour les cotes Rouen g226, g227 et g228 :

- Images consistant en des reproductions de microfilms : « Collections Bibliothèque municipale de Rouen ».
- Images consistant en des reproductions des manuscrits du Dictionnaire des idées reçues : « Collections Bibliothèque municipale de Rouen - photographie société Arkhénûm ».
- Autres images consistant en des reproductions des manuscrits de Bouvard et Pécuchet : « Collections Bibliothèque municipale de Rouen – photographie Thierry Ascencio-Parvy ».

Toute utilisation publique ou commerciale des images doit faire l'objet d'une autorisation préalable. Les demandes sont à adresser à la bibliothèque municipale de Rouen :
par courrier : Bibliothèque de Rouen, 3 rue Jacques Villon, F-76043 ROUEN CEDEX
ou par courriel : bibliotheque@rouen.fr

Pour la cote Montmorency : « Collection musée Jean-Jacques Rousseau - Ville de Montmorency - photographe Laure Querouil »
Toute utilisation publique ou commerciale des images doit faire l'objet d'une autorisation préalable. Les demandes sont à adresser au Musée Jean-Jacques Rousseau et Bibliothèque d'études rousseauistes par courrier :
Musée Jean-Jacques Rousseau et Bibliothèque d'études rousseauistes
4 rue du Mont-Louis
95160 Montmorency
ou par courriel : Rousseau-museum@ville-montmorency.fr

Pour la cote Antibes : « Collections privées »
Concernant l'utilisation des transcriptions :
L'utilisation des transcriptions à des fins privées, à des fins d'enseignement ou de recherche scientifique est autorisée, sous réserve de mentionner ainsi leur origine :

- « Transcription(s) réalisée(s) par [nom du transcripteur] pour l'édition des *Dossiers documentaires de Bouvard et Pécuchet*, sous la dir. de S. Dord-Crouslé, 2012-..., http://www.dossiers-flaubert.fr, ISSN 2495-9979. »

Pour toute publication, demander préalablement l'autorisation à la responsable de l'édition : Stéphanie Dord-Crouslé.
Le corpus complet est à citer comme suit :

- **Gustave Flaubert,** *Les dossiers documentaires de Bouvard et Pécuchet*. Édition intégrale balisée en XML-TEI accompagnée d'un outil de production de « seconds volumes » possibles, sous la dir. de Stéphanie Dord-Crouslé, 2012-..., http://www.dossiers-flaubert.fr, ISSN 2495-9979.

5. Responsabilités et ressources pour la gestion des données

- **Présentation de la section**

Cette section décrit, identifie, présente et nomme les responsables de la gestion des données.

- **Recommandations**

Afin de respecter les principes FAIR, CAHIER recommande le dépôt de celles-ci dans l'entrepôt Nakala (https://www.nakala.fr/). Ce service de dépôt et de stockage des données est proposé par la TGIR HumaNum pour les SHS. Il assure la gestion pérenne et sûre des données. Utiliser Nakala n'empêche pas de recourir à un second dépôt sur un autre entrepôt ou sur une plateforme institutionnelle.

Responsable de la gestion des données :

Dord-Crouslé, Stéphanie, IdHAL : stephanie-dord-crousle ; ORCID : 0000-0002-6683-9509, CNRS (UMR 5317 IHRIM), France

Évaluation des coûts (budgets, personnels et temps) dédiés à rendre les données FAIR (temps et budgets pour la collecte et la diffusion des données, pour le stockage et l'archivage).

Dans le cadre du Consortium CAHIER, les moyens assumés par l'infrastructure Huma-Num ont concerné les tâches suivantes:

- mise à disposition de moyens matériels tels que des serveurs, machines virtuelles, logiciels dédiés et licences supplémentaires dont les coûts et abonnements ne sont pas supportés par les projets, soit une économie estimée à ~5000€ / an pour chaque projet membre
- mise à disposition de moyens humains (ETP) pour des tâches spécifiques relevant à la fois de la gestion des moyens matériels (serveurs, machines, etc.), du stockage des données et des actions de formation, soit une économie estimée à plus de ~50000€ / an pour chaque projet membre

Equipe engagée dans la gestion des données à différentes étapes du projet (voir http://www.dossiers-flaubert.fr/projet-equipe-technique) :
Développements informatiques

- [2016-...] Pierre-Yves Jallud (CNRS-IHRIM)
- [2014-2015] Jean-Eudes Trouslard (jet@zoulous.com) pour le module "seconds volumes à la demande» :
  Parties plans et agencements de l'espace de travail.
  - module d'extraction des données de la base TEI vers la base mySql
  - conception et développement de l'affichage des agencements et plans d'une reconstitution conjecturale
  - outils de modifications de l'arborescence des agencements et plans
  - génération en pdf du texte de la reconstitution conjecturale
- [2011-2012] Hugo Schuler (CDD)
- [2009] Stéphane Wustner (Stagiaire)
- [2008-2009] Jérémie Lagravière (CDD)
- [2007-2008] Martial Tola (CNRS-ISH)
- [2007-2012, responsable] Raphaël Tournoy (CNRS-ISH)

TEI et ingénierie documentaire

- [2016-…] Maud Ingarao (ENS Lyon-IHRIM)
- [2016-…] Paul Gaillardon (CDD puis CNRS-IHRIM)
- [2016] Christelle Cluze (Stagiaire)
- [2012-...] Nathalie Arlin (Vacataire)
- [2011] Marjorie Burghart (EHESS, CIHAM)
- [2010-2015] Laetitia Faure (CDD puis CNRS-LIRE)
- [2008-2012, responsable] Emmanuelle Morlock-Gerstenkorn (CNRS-ISH)
- [2008] Vanessa Le Rolle (Stagiaire)
- [2007-2011] Christine Berthaud (CNRS-ISH)

Aide à la transcription

- [2011] Cécile Cordier (CDD)
- [2007-2008] Claire Giguet (CDD)

Traitement des images

- [2016-…] Florence Poncet (CDD IHRIM)
- [2011-2013] Françoise Notter-Truxa (CNRS-LIRE)
- [2007-2011] Véronique Églin (INSA, LIRIS)
- [2007-2011] Vincent Malleron (Doctorant, Université Lyon 2, LIRE et LIRIS)
- [2007-2010] Christophe Lemius (CNRS-LIRE)

6. Archivage des données

- **Présentation de la section**

Cette section décrit les données à conserver à court, moyen et long terme, les éventuelles données à détruire ou à laisser sous embargo et indique la durée de cette restriction.

- **Recommandations**

A l'issue du projet, des jeux de données se prêteront à une conservation à long terme pour une utilisation future, tandis que d'autres données ne nécessitent qu'une préservation à moyen terme car jugées moins essentielles et au potentiel de réutilisation limité, voire, elles pourront être destructibles pour des raisons de légalité ou de confidentialité.

Plateforme pour l'archivage pérenne des données

L'archivage n'est pas encore mis en place mais souhaité, et ce pour la globalité des données du projet. Les informations qui suivent sont donc de l'ordre du prospectif.
Néanmoins, grâce au dépôt prévu sur Nakala, Huma-Num devrait gérer l'archivage pérenne. Si cette solution n'est pas possible c'est un dépôt au CINES qui sera envisagé.

Durée de conservation des données

Illimitée

Volume des données à conserver

La totalité

Outils, méthodes, procédures nécessaires pour accéder à ces données archivées et les réutiliser

Voir conditions Huma-Num et CINES

7. Partage des données à l'issue / au fil du projet

- **Présentation de la section**

Cette section décrit la politique de dissémination des données. Elle indique s'il existe des limites à la diffusion des données, comment les données pourront être trouvées et réutilisées par les pairs, voire par le grand public.

- **Recommandations**

Une bonne dissémination des données requiert, dans la mesure du possible, le respect des principes FAIR : les données doivent être trouvables (findables), accessibles, interopérables et réutilisables. Pour être réutilisables, les données doivent être faciles d'accès, identifiables et citables grâce à des identifiants uniques (DOI) et leur usage facilité par l'accompagnement d'une description et de documentations, par des formats ouverts et non propriétaires et par une disponibilité facilitée par un lieu de stockage (entrepôt) ouvert, gratuit et référencé par les moteurs de recherche.

Les données primaires sont accessibles (images + transcriptions, certaines encore en cours). Les métadonnées du corpus sont partiellement moissonnables via OAI-PMH (voir https://www.rechercheisidore.fr/search/?source=10670/2.q6yiyl)

Eléments d'accompagnement qui permettent la réutilisation des données.

Des informations et tutoriels sont présents sur le site des *Dossiers*. Notamment sur la page "Espace de travail" qui renvoie à l'utilisation de l'Agenceur.

Publications sur les données destinées à en améliorer l'exposition

Le carnet de recherche dédié au projet de l'édition numérique des *Dossiers de Bouvard et Pécuchet* : https://flaubert.hypotheses.org/
Ce carnet diffuse les informations et actualités liées au projet et à son évolution, de même qu'il est un lieu d'échanges et de valorisation pour les chercheurs qui souhaitent réutiliser les sources des *Dossiers*.

L'inventaire des pièces du dossier de genèse de *Bouvard et Pécuchet* : https://flaubert.univ-rouen.fr/ressources/bp_sphere_inventaire.php

Bibliographie non exhaustive :

Stéphanie Dord-Crouslé. Vers une édition électronique des dossiers de Bouvard et Pécuchet. Stéphanie Dord-Crouslé, Stella Mangiapane et Rosa Maria Palermo Di Stefano. *Éditer le chantier documentaire de Bouvard et Pécuchet. Explorations critiques et premières réalisations numériques*, Andrea Lippolis Editore, pp.15-20, 2010. ⟨halshs-00549160⟩

Alexei Lavrentiev, Serge Heiden. Exploration textométrique du corpus des dossiers de Bouvard et Pécuchet. *Revue Flaubert,* Centre Flaubert, 2014, pp.1-12. ⟨halshs-00678874⟩

Stéphanie Dord-Crouslé. Le site et l'état d'avancement du projet Bouvard. Édition des dossiers documentaires de Bouvard et Pécuchet. *Journées d'études internationales des 11 et 12 décembre 2008, Lyon, École Normale Supérieure - Lettres et Sciences humaines*, Dec 2008, Lyon, France. ⟨halshs-00368846⟩

Pierre-Edouard Portier. Manipulations multimodales pour la construction de documents multistructurés. *Colloque: "Bouvard et Pécuchet : les " seconds volumes " possibles - Documentation, circulations, édition"*, ENS de Lyon, dir. Stéphanie Dord-Crouslé, Mar 2012, Lyon, France. ⟨halshs-00678876⟩

Caroline Angé. Édition de fragments : les enjeux de la mise en forme numérique. *colloque: "Bouvard et Pécuchet : les " seconds volumes " possibles - Documentation, circulations, édition"*, ENS de Lyon, dir. Stéphanie Dord-Crouslé, Mar 2012, Lyon, France. ⟨halshs-00678861⟩

Emmanuelle Morlock-Gerstenkorn. La pratique de l'encodage dans le projet d'édition électronique des Dossiers de Bouvard et Pécuchet : quelques exemples. Textes numériques : l'encodage, pratique savante ?, *Séminaire "Édition savante et humanités numériques"* (EHESS), Dec 2011, Paris, France. ⟨halshs-01141447⟩

Emmanuelle Morlock-Gerstenkorn. Les dossiers de Bouvard et Pécuchet de Flaubert - Fragments visuels et fragments logiques au sein du projet d'édition électronique. *Séminaire publication électronique* - IRHT Orléans, Dec 2009, Orléans, France. ⟨halshs-00438078⟩

Vincent Malleron. Outils d'analyse d'image pour les dossiers de Bouvard et Pécuchet : un panorama. *Édition des dossiers documentaires de Bouvard et Pécuchet. Journées d'études internationales des 11 et 12 décembre 2008, Lyon, École Normale Supérieure - Lettres et Sciences humaines*, Dec 2008, Lyon, France. ⟨halshs-00377381⟩

Stéphanie Dord-Crouslé. Fragments textuels et catégories de classement. Un cas d'utilisation de XML-TEI dans le dispositif éditorial du corpus BOUVARD. *Éditions critiques et génétiques en Rhône-Alpes*, Jun 2013, Grenoble, France. ⟨halshs-00838143⟩

Vincent Malleron. Le numérique et l'interdisciplinarité au service des dossiers de Bouvard et Pécuchet : Vers une mobilité retrouvée. *Séminaire de bilan et de prospective du Cluster 13 « Culture, patrimoine, création » mis en place et soutenu par la Région Rhône-Alpes*, vendredi 23 octobre 2009, Château de Montchat, Oct 2009, Lyon, France. ⟨halshs-00426391⟩

Stéphanie Dord-Crouslé, Emmanuelle Morlock-Gerstenkorn. Le "modèle abstrait" du corpus Bouvard : première approche. J*ournée d'étude " Constitution et exploitation de corpus issus de manuscrits - Lectures, écritures et nouvelles approches en recherche documentaire " organisée par Cécile Meynard et Thomas Lebarbé*, Mar 2009, Grenoble, France. ⟨halshs-00368044⟩

Vincent Malleron, Véronique Eglin, Hubert Emptoz, Stéphanie Dord-Crouslé, Philippe Régnier. *Hierarchical decomposition of handwritten manuscripts layouts. Computer Analysis of Images and Patterns*, Sep 2009, Muenster, Germany. pp.221-228, ⟨10.1007/978-3-642-03767-2⟩. ⟨halshs-00420059⟩

Stéphanie Dord-Crouslé, Emmanuelle Morlock, Raphaël Tournoy. **Nouveaux objets éditoriaux. Le site d'édition des dossiers documentaires de Bouvard et Pécuchet (Flaubert)**

Les Cahiers du numérique, Lavoisier, 2012, 7 (3-4/2011 " Empreintes de l'hypertexte. Rétrospective et évolution ", sous la dir. de Caroline Angé), pp.123-145. ⟨10.3166/LCN.7.3-4.123-145⟩

Vincent Malleron, Stéphanie Dord-Crouslé, Véronique Eglin, Hubert Emptoz, Philippe Régnier. Extraction automatisée de lignes et de fragments textuels dans les images de manuscrits d'auteur du XIXe siècle. *MAnifestation des JEunes Chercheurs en Sciences et Technologies de l'Information et de la Communication*, Nov 2009, Avignon, France. ⟨halshs-00443548⟩

Conditions de réutilisation : licences et contrats pour l'ensemble du projet

Conditions indiquées plus haut pour les images et manuscrits (voir : Accès, partage et limites d'accessibilité des données).

Pour les transcriptions : obligation de citation.

Schola Rhetorica

1. Plan de gestion de données (PGD) du projet SCHOLA RHETORICA

- **Présentation de la section**

Cette section décrit le PGD : elle présente l'auteur du PGD, les relecteurs du PGD, les autres intervenants assurant la gestion du PGD et, le cas échéant, ses mises à jour.

- **Recommandations**

Il est utile de désigner un responsable du PGD qui sera la personne à contacter. Il n'est pas nécessairement le responsable scientifique du projet. Il est recommandé d'associer ce responsable à son identifiant ORCID, IdRef, ISNI, IdHal et de nommer l'ensemble des personnes ayant contribué à la rédaction et à la relecture du PGD.
Le PGD évolue au fur et à mesure de l'avancée du projet de recherche et de l'enrichissement des données. Afin de faciliter sa rédaction, il est conseillé d'en produire une première version au début du projet, qui sera modifiée éventuellement en cours de projet, ainsi qu'à la fin du projet et d'indiquer les versions du PGD dans leur ordre antéchronologique en commençant par l'actuelle.

Auteures du plan de gestion des données :

NOILLE, Christine, ISNI : 0000000109073343 ; IdRef : https://www.idref.fr/032141025 professeure, Sorbonne Université, UMR 8599 CELLF, France
Rôle dans le projet : direction

L'HERMITE, Laurène, IdRef : https://www.idref.fr/236176927 ; Université de La Rochelle, Centre de recherches en histoire internationale et atlantique (EA1163), La Rochelle, France
Rôle dans le projet : co-auteure du PGD

Version du plan de gestion des données :

PGD SCHOLA-RHETORICA V1 : 30/10/2021

1 version de ce PGD est actuellement prévue

2. Présentation du projet et responsabilités

- **Présentation de la section**

Cette section décrit le projet ou le corpus sur lequel porte le PGD. Elle décrit le projet, ses objectifs, participants, etc. Ici, nous décrivons le Consortium CAHIER mais vous trouverez en annexe des exemples plus précis basés sur des projets membres de CAHIER

- **Recommandations**

Si le nom du projet est un acronyme, indiquez également la version développée.
Exemple : Antonomaz ("ANalyse auTOmatique et NumérisatiOn des MAZarinades")
Identifier la/le responsable scientifique du projet : Nom, Prénom, Institution, Laboratoire, Unité de rattachement, Ville, Pays. Mettre en lien son identifiant ORCID (ou ISNI, IdRef, IdHal, ...). Si possible indiquez des données de contact (courriel, téléphone professionnel)
Exemple : Karine ABIVEN (https://orcid.org/0000-0001-9518-1040), Sens-Texte-Informatique-Histoire (STIH), EA 4509, Université Paris - Sorbonne, Paris IV, France
Précisez également si le projet s'inscrit dans une programmation scientifique financée et les axes scientifiques liés à cette programmation :

- *Axes scientifiques d'un Labex*
- *Programme de financement d'un projet ANR, H2020*
- *Axe ou programme scientifique d'une structure de recherche liée au porteur ou à l'équipe projet...*

Nom du projet

SCHOLA RHETORICA

Responsable du projet (principal researcher) et unité de rattachement

NOILLE CHRISTINE, professeure, Sorbonne Université, UMR 8599 CELLF, France ; direction de SCHOLA-RHETORICA

Financeur(s) du projet et type de financement

2012 : ANR Hermès (porteur: Pr. F. Lavocat, Paris 7); 2014: Corpus CAHIER; 2012-2018: UMR 5316 Litt&Arts; 2018---- : UMR 8599 CELLF + UMR 5316 Litt&Arts

Institution / organisme / unité porteuses du projet

Unités co-porteuses : Sorbonne Université et Université Grenoble Alpes (convention de coportage en cours de signature)

Partenaires (identifier les organismes partenaires, ressources et co-financeurs du projet)

Partenaires institutionnels (d'après : https://schola-rhetorica.org/sr/A-propos/partenaires-et-credits)

- Université Grenoble Alpes
- Sorbonne Université
- CNRS
- UMR 5316 Litt&Arts
- UMR 8599 CELLF
- Labex OBVIL

- ANR
- TGIR Huma-Num / Consortium CAHIER

Descriptif et objectif(s) du projet

Schola Rhetorica, la recherche en rhétorique

De l'Antiquité au début du XXe siècle, la rhétorique s'élabore et s'enseigne par les manuels, les commentaires, les exercices, l'imitation. Le projet Schola Rhétorica rassemble des ressources éditoriales des XVIe-XIXe siècles (voir Présentation et description du corpus) pour approfondir l'ancienne rhétorique comme art de parler, comme art de lire et comme art d'écrire.

À l'école de la rhétorique

Trois types de ressources électroniques, trois axes de consultation du corpus :

- Définir les termes, avec le **GLOSSAIRE**
- Analyser les textes, avec les **COMMENTAIRES**
- Comprendre le système rhétorique, avec les **TRAITÉS**

Dates et durée

Date de début de financement et de début des travaux : 2012
Date de fin de financement et de fin des travaux : 2026

Mots clés du projet

- Rhétorique
- Humanisme
- Commentaires / Rhétorique – Commentaires de textes
- Glossaire
- Traités / Rhétorique – Traités

Publications (articles, pré-proposition, site web, …)

Site web du projet : https://schola-rhetorica.org/

Interface de consultation des textes : http://schola-rhetorica.fr/dev/#textes//langues/fr-fr-fr

Tous les articles des membres du projet, publiés sur la revue *Exercices de rhétorique* (https://journals.openedition.org/rhetorique/)

3. Présentation et description du corpus

- **Présentation de la section**

Cette section décrit le corpus et ses données. Elle décrit de façon plus précise les données du projet, les méthodes appliquées pour les collecter, etc. Ici, nous décrivons les données du Consortium CAHIER dans leur ensemble mais vous trouverez en annexe des exemples plus précis basés sur des projets membres de CAHIER

- **Recommandations**

Il s'agira de préciser le mode de collecte et l'origine des données, les centres d'archives, bibliothèques ou centres d'études hébergeant les données y compris si les données procèdent d'un moissonnage de ressources en ligne. L'organisation du corpus, l'arborescence des fichiers, le système de nommage et de gestion des répertoires et des fichiers doit être décrite. De même que la nature des données, leurs formats, leur volumétrie (en poids et nombre de fichiers), leur état, etc. Pour que les données soient réutilisables sur le long terme, les formats doivent être ouverts et non propriétaires et les données stockées dans des entrepôts accessibles.

Nom du projet

SCHOLA RHETORICA

Présenter et décrivez le corpus

Le corpus de Padoue

Le corpus de Padoue, que nous éditons ici dans la partie Commentaires est un ensemble très cohérent de quatre ouvrages publiés à Padoue de 1689 à 1729, chez le même éditeur, les Presses du Séminaire (ultérieurement imprimerie de Giovanni Manfrè), à savoir d'une part les trois commentaires rhétoriques de M. A. Ferrazzi, et d'autre part l'édition de la Rhétorique d'Aristote sur laquelle il s'appuie :

- 1689 : Aristote, De arte rhetorica, texte grec et en vis-à-vis la traduction latine de Marcantonio Majoragio (1514-1555), divisée en longs textus (paragraphes) numérotés
- 1694 : le commentaire de Ferrazzi sur 180 discours extraits de l'Histoire romaine de Tite-Live, *Exercitationes rhetoricae in orationes Titi Livii Patavini* (de nombreuses rééditions, par exemple dix rééd. de 1707 à 1710)
- 1694 : le commentaire de Ferrazzi sur 88 discours extraits de l'Énéide de Virgile, *Exercitationes rhetoricae in praecipuas P. Virgilii Maronis orationes, quae in Aeneidum libris leguntur* (de nombreuses rééditions, par exemple sept rééd. de 1720 à 1780, en Bavière)
- 1729 : le commentaire de Ferrazzi sur la totalité des 56 discours de Cicéron, *M. T. Ciceronis orationum cum argumentis, animadversionibus, et analysi M. Antonii Ferratii* (pas de rééd., mais l'éd. princeps est très répandue dans les bibliothèques)

Un corpus méthodique

Trois traits font de cet ensemble d'ouvrages édité par le Séminaire de Padoue un corpus méthodique.

- La circularité entre théorie et pratique : la Rhétorique d'Aristote est donnée avec un paragraphe propre à Padoue, et les analyses de discours que publie Ferrazzi renvoient uniquement à cette édition de la Rhétorique, avec une insistance toute particulière sur les passions du livre II (le pathos).
- La régularité : quoique très nombreuses, les analyses suivent constamment la même procédure et emploient toujours le même type de vocabulaire critique, dont elles accentuent la monotonie de façon délibérée, pédagogique.

- La masse : 180 discours tirés de Tite-Live, 88 de l'Énéide, les 56 discours de Cicéron, dont certains particulièrement longs. Une telle masse correspond aux nécessités internes de l'enseignement de la rhétorique. Elle permet de ramener à du sériel, donc à du reconnaissable, la diversité indéfinie et déroutante des situations rhétoriques concrètes. Et pour les étudiants d'aujourd'hui, elle remplace aussi le maître de l'époque, en signalant de façon récurrente quelles étaient à ses yeux les catégories vraiment importantes.

Corpus complémentaires

1. Une bibliothèque d'une vingtaine de traités de rhétorique des 17e-19e siècles : une dizaine est en cours de numérisation
2. Un glossaire collaboratif (indexant pour chaque terme des définitions attestées dans les traités depuis l'antiquité).

L'ensemble répondra aux standards d'un encodage TEI et d'une annotation enrichie (voir infra). Le point fort de cette plateforme numérique est non seulement de rassembler l'ensemble des corpus qui ont constitué l'empire de la rhétorique (intérêt patrimonial) mais d'offrir, grâce à l'élaboration d'interfaces dynamiques, une multitude de parcours pour une pluralité d'approches.

Période couverte par le corpus, auteur(s) concerné(s)

Période: XVIe-XIXe siècles. Auteurs: rhétoriciens de toute l'Europe

Organisation du corpus

Trois interfaces éditoriales : l'interface des traités, des commentaires, du glossaire.

Mode de collecte et origine des données

Données libres de droit ; OCRisation sous word ; relectures / corrections, puis encodage.

Etat du corpus numérique

- Pour les commentaires : ⅓ du corpus est en ligne (les commentaires sur Virgile)
- Pour les traités : une dizaine de traités sont édités
- Pour le glossaire collaboratif (wiki), plus de trois cents entrées sont multi-renseignées

Types de données :

Données textuelles

Volumétrie

- Commentaires : en l'état, 1 million de signes, à termes 3,5 millions
- Traités : en l'état, 7 millions de signes, à terme 12 millions
- Glossaire : en l'état 2 millions de signes, à terme 4 millions

En 2021, la volumétrie de l'ensemble des données est estimée à 3 Go.

Métadonnées, créées et standards et formats utilisés

Formats : MySQL,
Standards : TEI (sortie TEI prévue pour les Traités)

Les métadonnées descriptives, administratives et techniques

Pour les textes, glossaire et commentaires : Auteur / Titre français et latin / Edition française et latine / Sommaire / Séquençage.

Les métadonnées structurelles et l'annotation sémantique

Les traités ont fait l'objet d'un balisage sémantique au format interne avec la possibilité d'une transformation XML-TEI.

Référentiels d'indexation utilisés (vocabulaires contrôlés - thésaurus ou ontologies disciplinaires - et/ou indexation libre)

Indexation libre : références internes, y compris entre textes et termes rhétoriques.

4. Modalités de partage, de sauvegarde et de protection des données. Volumétrie des données stockées et espaces choisis.

- **Présentation de la section**

Cette section décrit la documentation produite au cours projet. Il s'agit d'une documentation autre que numérique (sur support papier par exemple). Si elle existe, il est important de la décrire. Cette section décrit également les lieux et infrastructures de stockage des données pendant le projet.

- **Recommandations**

Il s'agira de préciser ici le matériel physique et les lieux de stockage des données. Idéalement, il faudrait stocker les données dans au moins 2 endroits, éviter le stockage externe et privilégier les outils mis à disposition par l'institution. Pour cela, il peut être nécessaire de savoir quel est le volume approximatif des données à sauvegarder, l'espace de stockage nécessaire, la périodicité des sauvegardes, le nom et la nature du service fourni par l'institution, etc. On peut également indiquer les procédures de sauvegarde mises en place (fréquence des sauvegardes, automatisée ou non ?), les personnes en charge de la protection de ces données et du contrôle de l'accès, le mode de récupération des données en cas d'incident…

Accès, partage et limites des données

L'accès au site est libre sous licence Creative Commons BY-NC-SA.

Le **GLOSSAIRE** est un atelier numérique collaboratif (de type wiki), s'appuyant sur la seule exploitation de données référencées libres de droit.

Les **COMMENTAIRES** et les **TRAITÉS** : les textes sont numérisés à partir d'éditions libres de droit. Pour les **COMMENTAIRES**, les traductions sont de l'équipe éditoriale, sous licence Creative Commons BY-NC-SA.

5. Responsabilités et ressources pour la gestion des données

- **Présentation de la section**

Cette section décrit, identifie, présente et nomme les responsables de la gestion des données.

- **Recommandations**

Afin de respecter les principes FAIR, CAHIER recommande le dépôt de celles-ci dans l'entrepôt Nakala (https://www.nakala.fr/). Ce service de dépôt et de stockage des données est proposé par la TGIR HumaNum pour les SHS. Il assure la gestion pérenne et sûre des données. Utiliser Nakala n'empêche pas de recourir à un second dépôt sur un autre entrepôt ou sur une plateforme institutionnelle.

Responsable de la gestion des données

NOILLE, Christine, professeure, Sorbonne Université, UMR 8599 CELLF, France
Rôle dans le projet : Responsable du projet

Évaluation des coûts (budgets, personnels et temps) dédiés à rendre les données FAIR (temps et budgets pour la collecte et la diffusion des données, pour le stockage et l'archivage).

Dans le cadre du Consortium CAHIER, les moyens assumés par l'infrastructure Huma-Num ont concerné les tâches suivantes :

- mise à disposition de moyens matériels tels que des serveurs, machines virtuelles, logiciels dédiés et licences supplémentaires dont les coûts et abonnements ne sont pas supportés par les projets, soit une économie estimée à ~5000€ / an pour chaque projet membre
- mise à disposition de moyens humains (ETP) pour des tâches spécifiques relevant à la fois de la gestion des moyens matériels (serveurs, machines, etc.), du stockage des données et des actions de formation, soit une économie estimée à plus de ~50000€ / an pour chaque projet membre

Moyens humains : voir https://schola-rhetorica.org/sr/A-propos/l-equipe

6. Archivage des données

- **Présentation de la section**

Cette section décrit les données à conserver à court, moyen et long terme, les éventuelles données à détruire ou à laisser sous embargo et indique la durée de cette restriction.

- **Recommandations**

A l'issue du projet, des jeux de données se prêteront à une conservation à long terme pour une utilisation future, tandis que d'autres données ne nécessitent qu'une préservation à moyen terme car jugées moins essentielles et au potentiel de réutilisation limité, voire, elles pourront être destructibles pour des raisons de légalité ou de confidentialité.

La question de l'archivage des données sera intégrée au projet de développement de Schola Rhetorica déposé en 2022.

7. Partage des données à l'issue du projet

- **Présentation de la section**

Cette section décrit la politique de dissémination des données. Elle indique s'il existe des limites à la diffusion des données, comment les données pourront être trouvées et réutilisées par les pairs, voire par le grand public.

- **Recommandations**

Une bonne dissémination des données requiert, dans la mesure du possible, le respect des principes FAIR : les données doivent être trouvables (findables), accessibles, interopérables et réutilisables. Pour être réutilisables, les données doivent être faciles d'accès, identifiables et citables grâce à des identifiants uniques (DOI) et leur usage facilité par l'accompagnement d'une description et de documentations, par des formats ouverts et non propriétaires et par une disponibilité facilitée par un lieu de stockage (entrepôt) ouvert, gratuit et référencé par les moteurs de recherche.

La fairisation des données sera amorcée dans le projet 2022. L'accent sera mis sur la trouvabilité (nous disposons en général de bons systèmes de liens entre les textes) et l'accessibilité (on envisage de constituer un "usuel" en ligne, TLF, pour la rhétorique).

Publications sur les données destinées à en améliorer l'exposition

Dans une revue numérique (sur Openeditions), *Exercices de rhétorique*, créée et dirigée par la responsable de Schola-Rhetorica (Christine Noille) et son co-responsable pour la partie rhétorique (Francis Goyet)