Plan de gestion de données du projet ANR ObARDI

Victor Gay, Sébastien Poublanc, Jean-Luc Emmanuel Demonsant

DMP DU PROJET "OBARDI"

Plan de gestion de données créé à l'aide de DMP OPIDoR, basé sur le modèle "ANR - Modèle de PGD (français) - Personnalisé" fourni par Université Toulouse - Jean Jaurès.

RENSEIGNEMENTS SUR LE PLAN

Titre du plan	DMP du projet "ObARDI"
Version	Version initiale
Langue	fra
Date de création	2021-01-14
Date de dernière modification	2021-06-17
Identifiant	

RENSEIGNEMENTS SUR LE PROJET

Titre du projet ObARDI

Résumé

L'objectif d'ObARDI est d'améliorer notre compréhension des dynamiques de pouvoir qui sous-tendent la construction d'un État moderne en France. En effet, l'histoire des institutions d'Ancien Régime a été fortement marquée par un métarécit de la construction de l'État qui freine encore notre compréhension des mécanismes sous-tendant son développement (Blockmans et Genet, 1995 ; Blockmans et al., 2009). Malgré une ressemblance illusoire du vocabulaire, les catégories de l'État et de la société civile qui nous sont naturelles étaient sans doute étrangères aux sujets d'il y a trois siècles. De plus, nos représentations cognitives de ce qu'est un territoire politique restent limitées par un référentiel géographique qui ne s'est construit que progressivement au cours des dix-huitième et dix-neuvième siècles.

Pour dépasser ces difficultés, nous proposons un système innovant de représentation de la connaissance des dynamiques des institutions d'Ancien Régime à travers l'application d'ontologies aux données historiques : l'Infrastructure de Données d'Ancien Régime basée sur des Ontologies ["Ontology-based Ancien Régime Data Infrastructure" (ObARDI)]. Nous construisons cette infrastructure en quatre temps. Tout d'abord, nous assemblons une vaste base de données sur l'environnement local institutionnel, économique, et social de la France du dix-septième et dix-huitième siècles—une structure qui servira de matrice interopérable pour l'histoire d'Ancien Régime. Ensuite, nous intégrons ces données dans un système d'information pleinement structuré grâce à l'application d'ontologies formelles englobantes qui décrivent les liens entre les différentes entités de la base ainsi que les sources qui sous-tendent ce matériau. Après ce travail effectué, nous rendrons notre infrastructure conforme avec les principes de gestion des données FAIR dans la perspective des Données ouvertes liées : elle sera interopérable et accessible à travers une plateforme web ergonomique qui constituera un outil pour diffuser l'usage des humanités numériques envers différents publics. Enfin, cet environnement nous permettra de créer des outils cartographiques innovants afin de représenter les territoires d'Ancien Régime comme un feuilleté mouvant d'institutions dont les frontières furent structurées par des relations de pouvoir internes.

Comité : CE38 - Révolution numérique : rapports au savoir et à la culture

Aide de l'ANR: 612 920 euros

Début et durée du projet scientifique: janvier 2021 - 48 mois

Coordinateur du projet: Victor Gay (TSE, IAST, Fondation Jean Jacques Laffont, UT1)

Sources de financement	Agence Nationale de la Recherche (ANR) : ANR-20-CE38-0015

Partenaires	Maison de l'Homme et de la Société de Toulouse (201119593D)
	FONDATION JEAN JACQUES LAFFONT
	France, Amériques, Espagne, Sociétés, Pouvoirs, Acteurs (200311817L)
	Institut de Recherche en Informatique de Toulouse (199511949P)
	TEMPORA (201722259A)
	LITTORAL, ENVIRONNEMENT, TELEDETECTION, GEOMATIQUE (199612340K)

Produits de chercherche	ObARDI - Système d'information géographique (Jeu de données)
	ObARDI - Sources primaires (Collection)
	ObARDI - Plateforme OSF (Workflow)
	ObARDI - Ontologies (Modèle)
	ObARDI - Base de données (Jeu de données)

Contributeurs	Rôles
Victor Gay	Coordinateur du projet
	Personne contact pour les données (ObARDI - SIG, ObARDI - BD, ObARDI - SOURCES, ObARDI - ONTO, ObARDI - OSF)
Jean-Luc Demonsant	Responsable du plan de gestion de données

DMP DU PROJET "OBARDI"

1. DESCRIPTION DES DONNEES ET COLLECTE OU REUTILISATION DE DONNEES EXISTANTES

Le projet ObARDI produira quatre types de jeux de données (BD, SG, ONTO, SOURCES). Pour cela, elle recueillera des sources historiques primaires et réutilisera des données préexistantes : des sources historiques secondaires et des ontologies déjà construites. Le projet donnera aussi lieu à la publication du processus de recherche à partir de la plateforme créée sur l'infrastructure de l'Open Science Foundation (OSF).

Nouvelles données recueillies : sources historiques primaires

Pour créer l'infrastructure de données d'ObARDI, de nouvelles données seront recueillies à partir de sources historiques primaires. Celles-ci seront généralement disponibles sur Gallica, la bibliothèque numérique de la Bibliothèque nationale de France (BNF), Google Books, ou encore Archive.org. Si ce n'est pas le cas, nous procéderons à une numérisation des sources. Les formats de numérisation sont généralement le .pdf, ou le .jpg, ou le .tif pour les atlas historiques. A ce stade, la volumétrie globale est de 25 GB, mais celle-ci devra être mise à jour dans une version ultérieure du plan de gestion de données.

La liste ci-dessous est non-exhaustive et évoluera au fur et à mesure de l'avancement du projet.

- L'Atlas des Gabelles de Sanson (1665). Disponible sur Gallica.
- *Les États de dénombrement des ressorts des gabelles (1725-1726)*. Numérisé dans le cadre du projet à la BNF.
- Le Dénombrement du royaume de Saugrain (1709). Disponible sur Gallica.
- Le Nouveau dénombrement du royaume de Saugrain (1720). Disponible sur Gallica.
- Le Dictionnaire universelde Saugrain (1726). Disponible sur Gallica.
- Le Dictionnaire géographique, historique et politique des Gaules et de la France d'Expilly (1762-1770), en particulier, les appendices des tomes III et IV. Le tome 1 est disponible sur Google Books, les tomes 2, 3, 4, et 5, sur Archive.org, et le tome 6, sur Google Books.
- La Carte géographique des postes qui traversent la France de Sanson (1632). Disponible sur Gallica.
- Les *Livres de Poste*, publiés chaque année entre 1708 et la Révolution française. À ce stade, 28 volumes numérisés dans le cadre du projet à la BNF et à bibliothèque historique des postes et des télécommunications (BHPT).

Données préexistantes réutilisées

Nous mobiliserons deux types de données préexistantes : des sources historiques secondaires, et des ontologies déjà construites. Les listes ci-dessous sont non-exhaustives et évolueront au fur et à mesure de l'avancement du projet. Les formats de numérisation sont généralement le .pdf, ou le .jpg, ou le .tif pour les atlas historiques. A ce stade, la volumétrie globale est de 20 GB, mais celle-ci devra être mise à jour dans une version ultérieure du plan de gestion de données.

Sources historiques secondaires

Les 42 volumes de la collection Paroisses et communes de France coordonnés par le Laboratoire de Démographie Historique (LDH) sous la direction de Jean-Pierre Bardet et Claude Motte. L'ensemble de ces volumes a été numérisé dans le cadre du projet dans différentes bibliothèques universitaires.

Des atlas historiques nationaux et régionaux

- La Carte des généralités, subdélégations et élections en France à la veille de la Révolution de 1789 de Guy Arbellot, Guy, Jean-Pierre Goubert, Jacques Mallet, et Yvette Palazot (1986), publié aux Éditions du CNRS. Numérisé dans le cadre du projet.

- Le *Recueil de documents relatifs à la convocation des États généraux de 1789* d'Armand Brette (1894-1915*).* Disponible sur Archive.org (tome 1, 2, 3, 4).

- L'Atlas des bailliages ou juridictions assimilées ayant formé unité électorale en 1789 dressé d'après les actes de la convocation conservés aux Archives nationales d'Armand Brette (1904). Disponible sur Archive.org.

- L'Atlas historique : Provence, Comtat Venaissin, Principauté d'Orange, Comté de Nice, Principauté de Monaco d'Édouard Baratier, Georges Duby et Ernest Hildesheimer (1969), publié chez Armand Colin. Numérisé dans le cadre du projet.

- L'Atlas historique de la province de Languedoc coordonné par Élie Pélaquier (2009), publié par le Centre de Recherches Interdisciplinaires en Sciences humaines et Sociales (C.R.I.S.E.S) de l'Université Paul-Valéry Montpellier III. Disponible en version PDF sur le site du C.R.I.S.E.S.

- L'Atlas archéologique de Touraine coordonné par Elisabeth Zadora-Rio (2014), publié par la Revue Archéologique du Centre de la France (supplément au numéro 53). Disponible sous forme de publication électronique.

- L'Atlas historique du Limousin coordonné par le Centre de recherche interdisciplinaire en histoire, histoire de l'art et musicologie (CRIHAM), l'association "Rencontre des historiens du Limousin" et Éveha, un bureau d'études archéologique. Disponible sous forme de publication électronique.

- L'Atlas historique d'Alsace coordonné par le Centre de recherche sur les économies, les sociétés, les arts et les techniques de l'UHA. Disponible sous forme de publication électronique.

- L'Atlas historique de Lorraine coordonné par l'Université de Lorraine. Disponible sous forme de publication électronique.

- L'Atlas historique Auvergne, Bourbonnais, Velay coordonné par le centre d'histoire "Espaces et Cultures" de l'Université Clermont Auvergne. Disponible sous forme de publication électronique.

Des études régionales

- Les *Statistiques démographiques du Bassin Parisien (1636-1720)* de Jacques Dupâquier (1977), publié chez Gauthier-Villars. Numérisé dans le cadre du projet.

- *La population rurale du Bassin Parisien à l'époque de Louis XIV* de Jacques Dupâquier (1979), publié par les Éditions EHESS. Numérisé dans le cadre du projet.

- « Circonscriptions et régimes de l'impôt sur le sel de Normandie » de Jean-Marie Vallez (1982). Hors-série des *Annales de Normandie*. Recueil d'études offert en hommage au doyen Michel de Boüard Volume II. 1982.

Le fonds d'archive comprenant de informations sur 8 843 rébellions survenues entre 1661 et 1789, rassemblées dans les années 70, 80 et 90 par Jean Nicolas et des dizaines d'autres chercheurs. Voir La rébellion française. Mouvements populaires et conscience sociale (1661-1789)de Jean Nicolas (2002), publié chez Gallimard. Ce fonds d'archives unique se trouve désormais à la bibliothèque François-Lebrun de Rennes 2, dirigée par Renan Donnerh, partenaire de ce projet. Victor Gay et les partenaires du laboratoire TEMPORA, Gauthier Aubert, Philippe Hamon, Dominique Godineau, Renan Donnerh, ainsi que Johan Oszwald, partenaire du laboratoire LETG (Rennes 2), ont numérisé les 36 000pages de ce fonds.

Ontologies existantes

Un certain nombre d'ontologies existantes seront mobilisées au cours du projet :

- L'ontolgie symogih.org, un système modulaire de gestion de l'information historique

- L'ongologie micropublications, un modèle sémantique pour les propositions et arguments scientifiques.

- L'ontologie CIDOC-CRM, un modèle sémantique propre au patrimoine culturel.

- L'ontologie RiC-O, pour la description des documents d'archives et de leurs entités contextuelles.

- L'ontologie TSN, pour la description des nomenclatures territoriales statistiques.

Nouvelles données produites

Les nouvelles données produites dans le cadre du projet seront de quatre ordres :

- **ObARDI - BD** : la base de données de l'infrastructure ObARDI décrira pour chaque paroisse fiscale de 1789 quatre types d'informations de la seconde moitié du dix-septième siècle à 1789 : ses institutions, sa démographie, ses infrastructures de transport et ses rébellions. Seront rendus disponibles les processus de production de la base sous forme de scripts aux formats des logiciels Stata (fichiers .do) ou R (fichiers .R), les bases de données relationnelles créées dans le cadre du projet lors de l'extraction d'information à partir des sources primaires et secondaires aux formats des logiciels MS Access (fichiers .accdb) et FileMaker (fichiers .fmp12), ainsi que les données finales aux formats .csv (format libre) et .dta (format du logiciel Stata).

- **ObARDI - SIG** : le système d'information géographique créé à partir de la base de données. Seront rendus disponibles les processus de production sous forme de scripts au format python (fichiers .py) ainsi que les fonds de cartes finaux aux formats shapefiles (fichiers .shp, .shx, .dbf, .prj, .cpg) et GeoJSON (.geojson).

- **ObARDI - ONTO** : les ontologies structurant la base de données au format. owl.

- **ObARDI - SOURCES** : la collection des sources primaires et secondaires numérisées dans le cadre du projet.

- **ObARDI - OSF** : la plateforme OSF d'ObARDI, qui documente le *workflow* du projet.

La volumétrie globale des données produites reste à estimer et sera précisée dans une version ultérieure du plan de gestion de données, mais elle sera très probablement inférieure à 100 GB.

2. DOCUMENTATION ET QUALITE DES DONNEES

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

Chaque sous-ensemble des données partagées sera accompagné d'un fichier *README au* format .txt qui suivra les recommandations des Social Science Data Editors, mais aussi d'un data paper récrivant la méthodologie ayant présidé à la production des données, leur contexte institutionnel, ainsi qu'une description exhaustive des fichiers et des variables contenues dans le sous-ensemble. Nous suivrons le modèle définit par le journal Scientific Data du groupe *Nature*, qui est spécialisé dans la publication de *data papers*. Cette documentation sera aussi disponible en format html sur la plateforme finale d'ObARDI. Ces *data papers* seront disponibles sous licence CC BY 4.0 sur le portail HAL-ANR avec la mention de financement "ObARDI - ANR-20-CE38-0015 - AAPG2020 - 2020". Un exemple du type de *data paper* que nous produirons est celui réalisé par le porteur de projet, Victor Gay (2021) "Mapping the Third Republic. A Geographic Information System of France (1870-1940)" et qui paraîtra bientôt dans la revue *Historical Methods: A Journal of Quantitative and Interdisciplinary History*. Le document de travail est disponible sur HAL.

Des métadonnées suivant le vocabulaire Dublin Core accompagneront les données, ce qui permettra aussi à ces métadonnées d'être moissonnables via le protocole *Open Archives Initiative Protocol for Metadata Harvesting* (OAI-PMH).

La convention de nommage des fichiers et des dossiers sera celle préconisée par DORANum : des noms brefs et explicites, et pas de caractères spéciaux ni espaces. Les fichiers de travail intermédiaires (non diffusés) comporteront les dates de production et les numéros de version.

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

Différents contrôles de la qualité des données seront mis en place suivant le type de données. Ces contrôles seront décrits dans les *data papers* qui accompagneront chaque jeu de données, qui contiendront une section de "Validation technique" de la même manière que Gay (2021, pp. 23-5).

Par exemple, il sera procédé aux types de vérification suivants :

1. La saisie de données par des assistants de recherche sera contrôlée grâce aux fonctionnalités des formulaires MS Access, et un échantillonnage de leur travail pour double vérification de la qualité de la saisie sera réalisé systématiquement.

2. Les sources secondaires seront confrontées entre elles. Par exemple, le contenu des volumes Paroisses et Communes et France pour les variables "subdélégation" et "élections" en 1789 seront confrontées aux informations contenues dans la *Carte des généralités, subdélégations et élections en France à la veille de la Révolution de 1789 d'*Arbellot et al. (1986).

3. Des sous-ensembles des fichiers finaux seront confrontés à des travaux comparables. Par exemple, la variable "subdélégation" en 1789 de la base de données finale sera confrontée aux informations du fichier "Les subdélégations en France à la veille de la Révolution de 1789" produit par Cédric Chambru (2016) et disponible sur le site de partage de ressources géo-historiques du LARHRA.

4. Une confrontation des sources secondaires aux sources primaires sera systématiquement opérée via l'ontologie de critique de source développée dans le cadre du projet.

3. STOCKAGE ET SAUVEGARDE PENDANT LE PROCESSUS DE RECHERCHE

Tout au long du processus de recherche, les données de ce projet seront stockées sur le serveur institutionnel sécurisé du CNRS, MyCore, qui utilise le protocole sécurisé https. L'authentification se fait via la fédération d'identités CNRS Janus. Leur hébergement se fait sur des infrastructures dédiées, en France, dans le centre serveur du CNRS. En cas de perte de données, de vol ou de changement d'appareil, il suffit de « resynchroniser » l'appareil pour récupérer ses fichiers à J-1, J-6 ou J-15. Des copies des données sont également synchronisées sur chaque poste des membres du projet. L'accès aux données au cours du processus de recherche, stockées sur MyCore, est donné aux membres permanents du projet, et ponctuellement aux autres parties prenantes (stagiaires, etc.).

Le projet ne comporte pas de données personnelles à protéger.

4. EXIGENCES LEGALES ET ETHIQUES, CODES DE CONDUITE

Utilisation des œuvres sujettes au droit d'auteur

L'ensemble des œuvres sujettes au droit d'auteur utilisées dans le processus de construction de la base de données ObARDI rentre dans le cadre de l'exception de recherche telle que définie par l'article L. 122-5 du Code de la propriété intellectuelle. Cet article dispose que les auteurs des œuvres utilisées ne sont pas en mesure d'interdire "la reproduction d'une œuvre et sa représentation effectuées à des fins de conservation ou destinées à préserver les conditions de sa consultation à des fins de recherche ou d'études privées par des particuliers, dans les locaux de l'établissement ou sur des terminaux dédiés par des bibliothèques accessibles au public, par des musées ou par des services d'archives, sous réserve que ceux-ci ne recherchent aucun avantage économique ou commercial."

Dans la mesure où la finalité du projet ObARDI--projet de recherche public financé par l'ANR et ayant pour parties prenantes plusieurs institutions publiques de recherche--est de faciliter la consultation et l'utilisation de ressources à des fins de recherche, le projet rentre dans le cadre de l'exception de recherche.

Utilisation des œuvres sujettes au droit *sui generis* des bases de données

L'ensemble des œuvres sujettes au droit *sui generis* des bases de données utilisées dans le processus de construction de la base de données ObARDI rentre dans le cadre de l'exception de recherche telle que définie par l'article L. 342-3 5 du Code de la propriété intellectuelle. Cet article dispose que le producteur de la base de données ne peut interdire "les copies ou reproductions numériques de la base réalisées par une personne qui y a licitement accès, en vue de fouille de texte et de données incluses ou associées aux écrits scientifiques dans un cadre de recherche, à l'exclusion de toute finalité commerciale."

Dans la mesure où dans le cadre du projet ObARDI, il est question d'une recherche qui consiste à réunir un ensemble de sources, la démarche scientifique et de recherche du projet permet de légitimement penser qu'il rentre dans le cadre de l'exception de recherche.

Droits conférés aux producteurs et auteurs d'ObARDI

ObARDI est une base de données qui structure l'ensemble de son contenu de manière "originale" au sens du droit d'auteur. De plus, il s'agit d'un ensemble compilé issu d'un investissement substantiel des producteurs. Cette base de données est donc protégée par le droit *sui generis* des producteurs de bases de données. Elle est aussi protégée par un droit d'auteur en ce que les données sont organisées d'une manière originale, attestant d'un "effort intellectuel" des auteurs.

Dans la mesure où les articles L. 341-1 et L. 341-2 du code de la propriété intellectuelle relatif au droit *sui generis* des producteurs de bases de données ne mentionne pas l'existence d'une pluralité de producteurs, on se référera aux droits d'auteurs, qui eux spécifient ce cas de figure. ObARDI est issu de l'effort intellectuel d'une pluralité d'auteurs qui ont collectivement créé l'œuvre sous l'égide d'une autre personne—le chef de projet (Victor Gay)—et les apports de chaque auteur sont inidentifiables dans l'ensemble. Il s'agit donc d'une œuvre collective dont les droits d'auteur sont dévolus à la personne physique du chef de projet (Victor Gay) qui la divulgue.

Enfin, les données d'ObARDI ne sont pas des données à caractère personnel. Le protocole de recherche ne nécessite pas d'autorisation particulière relative aux questions éthiques. Le cas échéant, nous nous reporterons au guide "Pratiquer une recherche intègre et responsable" du COMETS (CNRS).

5. PARTAGE DES DONNEES ET CONSERVATION A LONG TERME

Les données issues du projet seront entreposées et diffusées sans embargo par l'ADISP-PROGEDO à l'issue du projet via le portail Quetelet-PROGEDO-Diffusion dans la rubrique "Données historiques" et seront mises à la disposition des chercheurs qui en feront la demande à travers le site dédié. Cet entrepôt de données possède les caractéristiques requises comme la gestion des métadonnées, la pérennité des formats via le CINES, un accès sécurisé, l'attribution d'un identifiant unique et pérenne (DOI) pour chaque jeu de données, etc. Il est en cours de certification par le Core Trust Seal (la liste des critères est disponible ici: https://doi.org/10.5281/zenodo.3638211). Afin de faciliter et encourager la réutilisation des bases de données issues du projet ObARDI, celles-ci seront diffusées sous "Licence Ouverte", licence compatible avec la licence CC BY 4.0. Les codes informatiques seront eux diffusés sous licence libre et *open source* MIT.

De plus, la plateforme interne de gestion du projet (plateforme ObARDI sur OSF) sera mise à disposition du public à la fin du projet pour permettre la documentation du processus de recherche lui-même.

L'accès aux données de la base ObARDI - DB pourra se faire avec n'importe quel éditeur de texte, tableur, ou logiciel statistique puisqu'il s'agira d'un format .csv pour les données BD. Il en est de même pour les ontologies de la base ObARDI - ONTO qui seront au format .owl. Les données au format .dta seront accessible par le logiciel Stata (néanmoins, toutes les données en .dta seront aussi disponibles en .csv). Les éléments de la collection ObARDI - SOURCES seront sous format image et PDF, donc il n'y a pas de logiciel spécifique nécessaire (mis à part Adobe Reader). Enfin, les fichiers de la base ObARDI - SIG seront accessibles par l'intermédiaire du logiciel libre QGIS (ou encore le logiciel propriétaire ArcGIS).

6. RESPONSABILITES ET RESSOURCES EN MATIERE DE GESTION DES DONNEES

Les responsables de la gestion des données sont :

- Jean-Luc Demonsant, IR PUD-T, MSHS-T (jean-luc.demonsant@univ-toulouse.fr)
- Victor Gay, TSE-IAST, porteur du projet (victor.gay@tse-fr.eu)
- Sébastien Poublanc, IR du projet ObARDI (sebastien.poublanc@univ-toulouse.fr)

Pour gérer les données et assurer leur FAIRisation, le projet ObARDI a prévu un poste ingénieur de recherche de la PUD-T financé par l'ANR à hauteur de 3 mois ETP. Jean-Luc Demonsant occupe ce poste.

Ce plan de gestion de données sera mis à jour annuellement tout au long du projet. Il s'agit de la version initiale.