

ÉCOLE NATIONALE DES CHARTES
UNIVERSITÉ PARIS, SCIENCES & LETTRES

Valentin De Craene

licencié en histoire

diplômé de master en histoire

agrégé d'histoire

Le Dictionnaire Numérique de la Ferme générale (1640-1794).

**Du traitement à la mise en valeur des
données nativement numériques.**

Mémoire pour le diplôme de master
« Technologies numériques appliquées à l'histoire »

2022

Résumé

Ce mémoire a été réalisé en vue de l'obtention du diplôme de Master 2 « Technologies numériques appliquées à l'histoire » de l'École Nationale des Chartes. Il a été rédigé à la suite d'un stage de quatre mois au sein de la Maison Européenne des Sciences de l'Homme et de la Société (Lille), dans le cadre de l'ANR FermeGé. Ce projet entend proposer une étude englobante de la Ferme générale (1640-1794) et de son emprise spatiale sur la société et les territoires d'Ancien Régime. Pour ce faire, un premier axe du projet s'attache à développer un dictionnaire historique regroupant une vaste collection de notices scientifiques. Ce mémoire étudie donc la chaîne de traitement des données scientifiques du dictionnaire, de la schématisation et de l'encodage automatisé jusqu'à la mise à disposition des données dans un prototype d'application web. Cette étude s'inscrit donc dans une approche critique des outils et méthodes mis en place dans le cadre du stage associé et du projet. Une réflexion sur la spécificité du dictionnaire comme objet scientifique structure ce travail.

Mots-clés : annotation sémantique ; Text Encoding Initiative (TEI) ; traitement des données ; encodage automatisé ; Python ; ODD ; Ferme générale ; fiscalité ; taxes ; époque moderne ; dictionnaire.

Informations bibliographiques : Valentin De Craene, *Le Dictionnaire numérique de la Ferme générale. Du traitement à la mise en valeur des données nativement numériques*, mémoire de master « Technologies numériques appliquées à l'histoire », dir. V. Le Fournier et E. Bermès, École nationale des chartes, 2022.

Remerciements

J'adresse mes plus sincères remerciements à l'ensemble des ingénieurs et chercheurs du pôle Humanités Numériques de la Maison Européenne des Sciences de l'Homme et de la Société et de l'ANR FermeGé qui m'ont accueilli au cours de ce stage. Mes pensées vont tout particulièrement à mesdames Victoria Le Fournier, directrice du stage, qui m'a encadré avec soin pendant ces quatre mois et Florence Perret, ingénieure en humanités numériques. Je remercie également monsieur Adrien Mével « co-stagiaire » pour le partage de son savoir-faire technique et son érudition en matière de *frontend*. Il me faut tout particulièrement saluer l'accueil chaleureux offert par madame Marie-Laure Legay, coordinatrice scientifique de l'ANR et messieurs Thomas Boullu et Benjamin Furst, responsables scientifiques du projet, avec qui j'ai eu le plaisir d'échanger sur les enjeux épistémologiques du projet.

Ce mémoire venant conclure l'année de master 2 Technologies Numériques Appliquées à l'Histoire, il me faut souligner le soutien sans failles de mes camarades de cette promotion 2021-2022. Au-delà de la franche camaraderie, j'ai trouvé au sein de l'École et du master un espace d'échange, d'enrichissement intellectuel et de débats sans commune mesure.

Mes remerciements vont de même à l'ensemble du corps professoral pour son érudition, sa bienveillance et sa disponibilité. Je remercie tout particulièrement monsieur Thibaut Clérice et madame Emmanuelle Bermés qui ont participé à l'encadrement de ce présent mémoire. J'adresse ma sincère gratitude à monsieur Olivier Poncet pour avoir accepté de présider le jury de ce présent travail.

Ce tableau ne pourrait être complet sans remercier ma compagne, Olivia, pour son soutien permanent.

Bibliographie

Développement applicatif et web

BARDIOT (Clarisso), *Happy APIs : Débridons les APIS pour développer les humanités numériques*, DORRA-DH, 2018, URL : <https://dorradh.hypotheses.org/66> (visité le 19/07/2022).

FIELDING (Roy T), *Architectural styles and the design of network-based software architectures*, OCLC : 45706361, University of California, 2000.

GRINBERG (Miguel), *Flask Web Development : Developing Web Applications with Python*, seconde édition, 2018.

LE FOURNER (Victoria) et CHAGUÉ (Alix), « Structuration automatique du texte des jugements du Conseil des Tissus parisiens (ANR TIME US) », (, 17 févr. 2022), DOI : 10.34847/nk1.853980ck.

SCHULTZ (Emilien) et BUSSONNIER (Matthias), *Python pour les SHS : introduction à la programmation pour le traitement de données*, ISSN : 2269-4714, Rennes, France, 2021.

Encodage et gestion des données

BISSON (Marie), KUHRY (Emmanuelle) et GOLOUBKOFF (Anne), *Editer un inventaire ancien en XML-TEI P5*, 2019, URL : hal-02457987.

BOHBOT (Hervé), FRONTINI (Francesca), LUXARDO (Giancarlo), KHEMAKHEM (Mohamed) et ROMARY (Laurent), « Presenting the Nénufar Project : a Diachronic Digital Edition of the Petit Larousse Illustré », dans *GLOBALEX 2018 - Globalex workshop at LREC2018*, Miyazaki, Japan, 2018, p. 1-6, URL : <https://hal.archives-ouvertes.fr/hal-01728328> (visité le 12/04/2022).

BUDIN (Gerhard), MAJEWSKI (Stefan) et MÖRTH (Karlheinz), « Creating Lexical Resources in TEI P5 », *Journal of the Text Encoding Initiative* (, 5 nov. 2012), Number : Issue 3 Publisher : Text Encoding Initiative Consortium, DOI : 10.4000/jtei.522.

BURNARD (Lou), *Qu'est-ce que la Text Encoding Initiative ?*, Marseille, France, 2015, URL : <http://books.openedition.org/oep/1237> (visité le 15/08/2022).

- BURNARD (Lou), *ODD Chaining for Beginners*, URL : <http://teic.github.io/TCW/howtoChain.html> (visité le 28/06/2022).
- CONSORTIUM (TEI), *TEI P5 : guidelines for Electronic Text Encoding and Interchange*, dir. Syd Bauman et Lou Burnard, 2 t., Oxford, Royaume-Uni de Grande-Bretagne et d'Irlande du Nord, 2008.
- CORBIÈRES (Caroline), *Du catalogue au fichier TEI, création d'un workflow pour encoder automatiquement en XML-TEI des catalogues d'exposition*, Mémoire de master TNAH, Paris, Ecole Nationale des Chartes, 2020.
- GROVER (Claire), GIVON (Sharon), TOBIN (Richard) et BALL (Julian), « Named Entity Recognition for Digitised Historical Texts », *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2008, 26 May - 1 June 2008, Marrakech, Morocco* (, 2008), p. 1343-1346, URL : <https://www.research.ed.ac.uk/en/publications/named-entity-recognition-for-digitised-historical-texts> (visité le 02/05/2022).
- IDE (Nancy), ET (Ab) et VÉRONIS (Jean), « Codage TEI des dictionnaires électroniques », *Cahiers GUTenberg* (, 1^{er} janv. 1996), DOI : 10.5802/cg.197.
- JANÈS (Juliette), *Du catalogue papier au numérique Une chaîne de traitement ouverte pour l'extraction d'informations issues de documents structurés*, Mémoire de master TNAH, Paris, Ecole nationale des chartes, 2021, URL : https://raw.githubusercontent.com/Juliettejns/Memoire_TNAH/main/Jjanès_Mémoire.pdf (visité le 29/05/2022).
- KHEMAKHEM (Mohamed), FOPPIANO (Luca) et ROMARY (Laurent), « Automatic Extraction of TEI Structures in Digitized Lexical Resources using Conditional Random Fields », dans *electronic lexicography, eLex 2017*, Leiden, Netherlands, 2017, URL : <https://hal.archives-ouvertes.fr/hal-01508868> (visité le 12/04/2022).
- KHEMAKHEM (Mohamed), GABAY (Simon), JOYEUX-PRUNEL (Béatrice), ROMARY (Laurent), SAINT-RAYMOND (Léa) et RONDEAU DU NOYER (Lucie), « Information Extraction Workflow for Digitised Entry-based Documents », dans *DARIAH Annual event 2020*, Zagreb / Virtual, Croatia, 2020, URL : <https://hal.archives-ouvertes.fr/hal-02508549> (visité le 12/04/2022).
- LE FOURNER (Victoria), *Étude de la structuration automatique et de l'éditorialisation d'un corpus hétérogène, l'exemple des sources du conseil des prud'hommes pour le textile du XIX^e siècle*. Mémoire de master TNAH, Paris, 2019.
- MANGEOT (Mathieu) et ENGUEHARD (Chantal), « Des dictionnaires éditoriaux aux représentations XML standardisées », *Nuria Gala et Michael Zock. Ressources Lexicales : contenu, construction, utilisation, évaluation* (, 2013), Publisher : John Benjamins, p. 24, DOI : 10.1075/lis.30.08man.
- MÖRTH (Karlheinz), ROMARY (Laurent), BUDIN (Gerhard) et SCHOPPER (Daniel), « Modeling Frequency Data : Methodological Considerations on the Relationship between Dictionaries and Corpora », *Journal of the Text Encoding Initiative* (, 28 déc.

- 2014), Number : Issue 8 Publisher : Text Encoding Initiative Consortium, DOI : 10.4000/jtei.1356.
- NAGAI (Noriyoshi), KIMURA (Fuminori), MAEDA (Akira) et AKAMA (Ryo), « Personal Name Extraction from Japanese Historical Documents Using Machine Learning », dans *International Conference on Culture and Computing*, 2015, p. 207-208, DOI : 10.1109/Culture.and.Computing.2015.46.
- RONDEAU DU NOYER (Lucie), *Encoder automatiquement des catalogues en XML/TEI, principes, évaluations et application à la revue des autographes de la librairie Charavay*, Mémoire de master TNAH, Paris, Ecole Nationale des Chartes, 2019.
- SCHWARTZ (Daniel L.), GIBSON (Nathan P.) et TORABI (Katayoun), « Modeling a Born-Digital Factoid Prosopography using the TEI and Linked Data », *Journal of the Text Encoding Initiative* (, 21 mars 2022), Publisher : Text Encoding Initiative Consortium, DOI : 10.4000/jtei.3979.
- ZUNKE (Saurabh) et D'SOUZA (Veronica), « JSON vs XML : A Comparative Performance Analysis of Data Exchange Formats », *International Journal of Computer Science and Network*, 3–4 (2014), p. 5.

Entités nommées et TAL

- BORIN (L.), KOKKINAKIS (D.) et OLSSON (Leif-Jöran), « Naming the Past : Named Entity and Animacy Recognition in 19th Century Swedish Literature », dans *La-TeCH@ACL*, 2007.
- BYRNE (Kate), « Nested Named Entity Recognition in Historical Archive Text », dans *School of Informatics*, University of Edinburgh, 2007, p. 589-596, DOI : 10.1109/ICSC.2007.107.
- CRANE (Gregory) et JONES (Alison), « The challenge of virginia banks : an evaluation of named entity analysis in a 19th-century newspaper collection », dans *Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries*, New York, NY, USA, 2006 (JCDL '06), p. 31-40, DOI : 10.1145/1141753.1141759.
- EHRMANN (Maud), *Les entités nommées, de la linguistique au TAL : statut théorique et méthodes de désambiguisation*, These de doctorat, Paris 7, 2008, URL : <http://www.theses.fr/2008PA070095> (visité le 14/04/2022).
- GALLERON (Ioana), PATRAŞ (Roxana), GRĂDINARU (Camelia) et MÉLANIE-BECQUET (Frédérique), « À la recherche des haïdouks. Annoter les entités spatiales dans des romans roumains du xixe siècle », *Humanités numériques*–3 (1^{er} mai 2021), Number : 3 Publisher : Humanistica, DOI : 10.4000/revuehn.1399.
- JURAFSKY (Daniel) et MARTIN (James H.), *Speech and Language Processing. An introduction to Natural Language Computational Linguistics, and Speech Recognition*,

Third Edition draft, 2022, URL : https://web.stanford.edu/~jurafsky/slp3/ed3book_jan122022.pdf.

KHAN (Fahad), ROMARY (Laurent), SALGADO (Ana), BOWERS (Jack), KHEMAKHEM (Mohamed) et TASOVAC (Toma), « Modelling Etymology in LMF/TEI : The Grande Dicionário Houaiss da Língua Portuguesa Dictionary as a Use Case », dans *LREC 2020 - 12th Language Resources and Evaluation Conference*, Marseille, France, 2020, URL : <https://hal.inria.fr/hal-02618067> (visité le 12/04/2022).

NOUVEL (Damien), *Reconnaissance des entités nommées par exploration de règles d'annotation : interpréter les marqueurs d'annotation comme instructions de structuration locale*. These de doctorat, Tours, 2012, URL : <http://www.theses.fr/2012TOUR4011> (visité le 14/04/2022).

OMRANE (Nouha), NAZARENKO (Adeline) et SZULMAN (Sylvie), « Les entités nommées : éléments pour la conceptualisation », dans *21es Journées francophones d'Ingénierie des Connaissances*, Nîmes, 2010, URL : <https://hal.archives-ouvertes.fr/hal-00525530> (visité le 25/03/2022).

WON (Miguel), MURRIETA-FLORES (Patricia) et MARTINS (Bruno), « Ensemble Named Entity Recognition (NER) : Evaluating NER Tools in the Identification of Place Names in Historical Corpora », *Frontiers in Digital Humanities*, 5 (2018), URL : <https://www.frontiersin.org/articles/10.3389/fdigh.2018.00002> (visité le 16/08/2022).

Histoire et historiographie de la Ferme générale

AZIMI (Vida), *Un modèle administratif de l'Ancien régime : les commis de la Ferme générale et de la Régie générale des aides*, Paris, France, 1987.

BARBICHE (Bernard), *Les institutions de la monarchie française à l'époque moderne : XVIe-XVIIIe siècle*, ISSN : 1630-5264, Paris, France, 1999.

BÉAUR (Gérard), BONIN (Hubert) et LEMERCIER (Claire), *Fraude, contrefaçon et contrebande, de l'Antiquité à nos jours* : ISSN : 1422-7630, Genève, Suisse, 2006.

BEAUREPAIRE (Pierre-Yves) et CORNETTE (Joël), *La France des Lumières*, Paris, France, 2014.

DESSERT (Daniel), *L'argent du sel : le sel de l'argent*, Paris, France, 2013.

DRÉVILLON (Hervé) et CORNETTE (Joël), *Les rois absolus*, Paris, France, 2014.

DURAND (Yves), *Les fermiers généraux au XVIIIe siècle*, ISSN : 0078-9895, Paris, France, 1971.

GESLOT (Jean-Charles) et REBOLLEDO-DHUIN (Viera), « Du livre imprimé au Web sémantique : le projet du Dictionnaire des éditeurs français du xixe siècle », *Humanités numériques-2* (1^{er} juin 2020), DOI : 10.4000/revuehn.426.

- KWASS (Michael) et TAFFIN-JOUHAUD (Dominique), *Louis Mandrin : la mondialisation de la contrebande au siècle des Lumières*, ISSN : 2260-7552, Paris, France, 2016.
- LEGAY (Marie-Laure), « Défier l'administration : les inscriptions en faux contre la Ferme générale (1680-1780) », *Revue historique*, 690-2 (2019), ISBN : 9782130803188 Publisher : Presses Universitaires de France, p. 315-334, DOI : 10.3917/rhis.192.0315.
- LEGAY (Marie-Laure), DUBET (Anne), FÉLIX (Joël) et HOCQUET (Jean-Claude), *Dictionnaire historique de la comptabilité publique : vers 1500-vers 1850*, Rennes, France, 2010.
- LEGAY (Marie-Laure) et GUINET (Philippe Préfacier), *Les États provinciaux dans la construction de l'État moderne aux XVIIe et XVIIIe siècles*, ISSN : 1420-7699 Type : Texte remanié de, Genève, Suisse, 2001.
- LOGETTE (Aline), « La Régie générale au temps de Necker et de ses successeurs (1777-1786) », *Revue historique de droit français et étranger* (1922-), 60-3 (1982), Publisher : Editions Dalloz, p. 415-445, URL : <https://www.jstor.org/stable/43846745> (visité le 02/05/2022).
- MARKOVIC (Momcilo), SERNA (Pierre Préfacier) et COLLIN (Bruno Préfacier), *Paris brûle !: l'incendie des barrières de l'octroi en juillet 1789*, ISSN : 2274-3804 Type : Texte remanié de, Paris, France, 2019.
- MATTHEWS (George T.), *The royal general farms in eighteenth-century France*, New York, Etats-Unis d'Amérique, 1958.
- MOUSNIER (Roland), *Les institutions de la France sous la monarchie absolue : 1598-1789*, 2 t., Paris, France, 1974.
- *La Fonction publique en France du début du seizième siècle à la fin du dix-huitième siècle*, Paris, France, 1976.
- NICOLAS (Jean), *La rébellion française : mouvements populaires et conscience sociale*, ISSN : 0083-3673, Paris, France, 2002.
- PEYSSON (Jean-Marc), *Le mur d'enceinte des fermiers-généraux (1784-1791) : politique, économie, urbanisme*, Number Of Volumes : 2, Thèse de 3e cycle, France, 1983.
- ROBISHEAUX (Earl E.), « The "Private Army" of the Tax Farms : The Men and their Origins », *Histoire sociale / Social History*, 6-12 (1973), Number : 12, URL : <https://hssh.journals.yorku.ca/index.php/hssh/article/view/40719> (visité le 02/05/2022).

Humanités numériques et science ouverte

- ACQUIER (Françoise), *Quelles images puis-je publier dans ma thèse ? Réponse en pratique dans un atelier pédagogique*, Ethique et droit, URL : <https://ethiquedroit.hypotheses.org/2947> (visité le 27/07/2022).

- BOULAIRE (Cécile) et CARABELLI (Romeo), « Chapitre 7. Du digital naïve au bricoleur numérique : les images et le logiciel Omeka », dans *Expérimenter les humanités numériques : Des outils individuels aux projets collectifs*, dir. Étienne Cavalié, Frédéric Clavert, Olivier Legendre et Dana Martin, Montréal, 2018 (Parcours numérique), p. 81-103, URL : <http://books.openedition.org/pum/11115> (visité le 01/08/2022).
- CAVALIÉ (Étienne), *L'indexation matière en transition : de la réforme de Rameau à l'indexation automatique*, ISSN : 0184-0886, Paris, France, 2019.
- CHARLES (Louise-Anne), *Retour d'expériences : publication de données dans l'entrepôt de données SHS Nakala*, Lab & doc, URL : <https://labedoc.hypotheses.org/10424> (visité le 01/08/2022).
- CLAVERT (Frédéric), DANIEL (Johanna), FLECKINGER (Hélène), GRANDJEAN (Martin) et IDMHAND (Fatiha), « Histoire et humanités numériques : nouveaux terrains de dialogue entre les archives et la recherche », *La Gazette des archives*, 245–1 (2017), p. 121-134, DOI : [10.3406/gazar.2017.5519](https://doi.org/10.3406/gazar.2017.5519).
- CRYMBLE (Adam), *Technology and the historian : transformations in the digital age*, Urbana, Etats-Unis d'Amérique, 2021.
- DACOS (Marin), « Des nains sur les épaules de géants : ouvrir la science en France », *Revue Politique et Parlementaire* (, 2019), Publisher : Colin, URL : <https://hal.archives-ouvertes.fr/hal-02366604> (visité le 18/07/2022).
- GESLOT (Jean-Charles) et REBOLLEDO-DHUIN (Viera), « Du livre imprimé au Web sémantique : le projet du Dictionnaire des éditeurs français du xixe siècle », *Humanités numériques*–2 (1^{er} juin 2020), DOI : [10.4000/revuehn.426](https://doi.org/10.4000/revuehn.426).
- GUICHARD (Éric), *Les humanités numériques n'existent pas*, août 2019, URL : <https://hal.archives-ouvertes.fr/hal-02403315> (visité le 22/06/2022).
- JOYEUX-PRUNEL (Béatrice), « Bases de données et gestion de projets en humanités numériques », *Biens Symboliques / Symbolic Goods. Revue de sciences sociales sur les arts, la culture et les idées*–2 (19 févr. 2018), Number : 2 Publisher : Presses Universitaires de Vincennes, DOI : [10.4000/bssg.242](https://doi.org/10.4000/bssg.242).
- KUNSTMANN (Pierre), GERNER (Hiltrud) et SOUVAY (Gilles), « Le Dictionnaire Électro-nique de Chrétien de Troyes (DÉCT1) : révision et élargissement », dans *XXVIème Congrès International de Linguistique et de Philologie Romane*, Valencia, Spain, 2010, à paraître, URL : <https://hal.archives-ouvertes.fr/hal-00522494> (visité le 12/04/2022).
- Stéphane Lamassé, Gaétan Bonnot et Laboratoire de médiévistique occidentale de Paris (éd.), *Dans les dédales du web : historiens en territoires numériques*, OCLC : 1104307657, 2019.

- REBOUILLAT (Violaine), *Ouverture des données de la recherche : de la vision politique aux pratiques des chercheurs*, These de doctorat, Paris, CNAM, 2019, URL : <https://www.theses.fr/2019CNAM1254> (visité le 28/06/2022).
- REHR (Jean-Paul) et BEAULIEU (Marie-Anne Polo de), « Thesaurus Exemplorum Medii Aevi : une base de données collaborative sur les exempla médiévaux », *Humanités numériques*-4 (1^{er} déc. 2021), DOI : 10.4000/revuehn.2630.
- RICARD, *Le nouveau régime juridique de la réutilisation des informations publiques*, Droit(s) des archives, URL : <https://siafdroit.hypotheses.org/659> (visité le 27/07/2022).
- SCHÖPFEL (Joachim), FARACE (Dominic), PROST (Hélène) et ZANE (Antonella), « Data Papers as a New Form of Knowledge Organization in the Field of Research Data », *KNOWLEDGE ORGANIZATION*, 46-8 (2019), p. 622-638, DOI : 10.5771/0943-7444-2019-8-622.
- SUPÉRIEUR ET DE LA RECHERCHE (Ministère de l'Enseignement), *Plan national pour la science ouverte*, 2018, URL : https://www.enseignementsup-recherche.gouv.fr/sites/default/files/content_migration/document/PLAN_NATIONAL_SCIENCE_OUVERTE_978672.pdf (visité le 28/06/2022).
- *Deuxième plan national pour la science ouverte 2021-2024*, 2021, URL : https://www.ouvrirlascience.fr/wp-content/uploads/2021/06/Deuxieme-Plan-National-Science-Ouverte_2021-2024.pdf (visité le 18/07/2022).
- THIBAUT (Françoise), *Mutations des sciences humaines et sociales : Les Maisons des Sciences de l'Homme et leur réseau*, dir. Réseau national des maisons des sciences de l'homme et Alliance Athena, Paris, France, 2021.
- WAQUET (Françoise), *Dans les coulisses de la science : techniciens, petites mains et autres travailleurs invisibles*, Paris, France, 2022.

Introduction

Les « humanités numériques » n'existent pas, ni en tant que discipline ni en tant que champ de savoir. Ses partisans se gardent d'ailleurs bien de les définir, préférant parler de la « construction d'un milieu » [Berra, 2015]. Ce serait peut-être un syndicat. Ce constat n'empêche pas que des personnes se regroupant sous cette bannière réalisent des travaux exceptionnels. Scientifiquement, l'enjeu est celui des « méthodes numériques (ou digitales) pour sciences sociales », et le point important est celui de l'appropriation et de la participation au façonnage de la culture de l'écrit contemporain au sein de chaque discipline (actuelle ou à venir)¹.

Telle est la conclusion tirée par Eric Guichard de son analyse des humanités numériques et de leur place dans le champ des disciplines académiques. Le philosophe préfère déplacer la focale des méthodes des humanités numériques vers la question de la construction d'une nouvelle forme de « culture numérique de l'écrit contemporain² », à savoir une interaction nouvelle entre l'aspect technique des nouvelles technologies et le versant intellectuel de la recherche. S'il ne nous appartient pas de prendre position dans ce débat sur l'existence ou non des humanités numériques comme champ disciplinaire autonome, force est de constater que ce dialogue entre le développement et la maîtrise des outils numériques d'une part, et la recherche scientifique académique d'autre part, se trouve au cœur de nombreux projets de recherche relatifs aux humanités et à l'histoire. A cet égard, le projet ANR FermeGé s'inscrit à plusieurs échelles dans un dialogue entre technique et recherche, participant d'une nouvelle forme de l'écrit numérique mentionnée par Eric Guichard. Ce projet, visant à comprendre l'emprise de la Ferme générale (1640-1794) sur une mosaïque de territoires et de groupes sociaux d'Ancien Régime, se base sur un constat :

Tous les historiens conviennent de l'importance de la Ferme générale dans le paysage financier et politique de la France d'Ancien régime, et tous déplorent le caractère partiel des connaissances sur cette grande compagnie.

Tel est le bilan dressé par les historiens Thomas Boullu et Marie-Laure Legay dans la notice liminaire du carnet Hypothèses.org introduisant le projet de constitution d'un dictionnaire historique de la Ferme générale, en mai 2019³. Ce manque historiographique ne peut que surprendre l'historien contemporain qui s'intéresse à cette institution financière en charge de la perception des aides entre 1640 et 1794. Effectivement, si la Ferme générale a laissé une postérité dans les représentations populaires de la seconde modernité, autour des figures de ses agents, des « gabeleurs » ou des contrebandiers l'ayant contestée tel que Mandrin, la production scientifique sur ce sujet reste relativement peu étayée. Dès lors, c'est ce vide historiographique, notamment à l'échelle locale, que l'A.N.R FermeGé

1. Éric Guichard, *Les humanités numériques n'existent pas*, août 2019, URL : <https://hal.archives-ouvertes.fr/hal-02403315> (visité le 22/06/2022), p.10.

2. *Ibid.*, p.11.

3. <https://dicofg.hypotheses.org/category/presentation-du-dictionnaire>.

entend combler, par la mise en oeuvre d'un dictionnaire relevant de l'histoire « totale » et croisant les échelles comme les thématiques d'analyses. La problématique formulée dans le cadre globale du projet est donc la suivante :

La Ferme générale, dénuée des attributs symboliques du pouvoir, mais dotée de moyens de coercition (jugés importants jusqu'à présent), a-t-elle figé un mode de gouvernance arbitraire du territoire et de la société, compris à la fois comme mode rationnel de gestion, et comme processus de civilisation ?

Afin de saisir l'ampleur de l'emprise de la Ferme générale et de « l'administration du privilège » à l'oeuvre durant cette période, l'analyse se veut exhaustive. C'est pourquoi la réalisation d'un vaste et dense corpus de notices structurées au sein d'un dictionnaire nativement numérique, et adossé à un atlas, s'est imposé comme un support technique pertinent. Dans un premier temps, les notices furent déposées sur le carnet Hypothèses.org du projet FermeGé⁴. Cependant, la décision fut prise d'arrêter ce dépôt sur la plateforme en raison des impératifs de pérennisation des données dans le cadre de l'ANR. A cette même occasion, il fallut proposer un adaptation du format de données des notices permettant une plus vaste ouverture et réutilisation au sein de la communauté scientifique concernée. La proposition d'encoder les notices et la structure générale du dictionnaire en XML/TEI s'est avérée être une alternative au carnet Hypothèses.org, permettant d'ouvrir et pérenniser les données d'une part, et d'autre part de transformer ces fichiers pour faciliter l'intégration des données scientifiques dans un site internet dédié.

C'est dans ce contexte de transition technique que s'est inscrit le stage dont ce mémoire est issu. La nécessité de mettre en oeuvre un schéma d'encodage des notices et sa documentation, puis un processus d'automatisation de la pose des balises et enfin la mise en oeuvre d'un prototype reflétant les possibilités techniques du produit final se sont imposés à nous comme des étapes logiques et structurantes d'un *Workflow* de traitement de la donnée. Cependant, il nous semble nécessaire de définir au préalable les concepts au coeur de cette réflexion et de nos réalisations techniques.

En premier lieu, comme nous l'avons suggéré, l'encodage des notices a été réalisé en XML/TEI, à savoir un langage d'encodage formel, structuré par des balises sémantiques et se conformant aux recommandations du Consortium TEI⁵. La *Text Encoding Initiative* (TEI) propose donc un cadre d'encodage et réflexion sur la nature des documents textuels, puisque l'apport du sémantisme au texte brut en est l'objectif premier. Depuis 2001 et la 4^e version des *Guidelines*, la TEI s'appuie sur l'*eXtended Markup Language* (XML), un métalangage de balisage, permettant une plus grande interopérabilité dans un

4. <https://dicofg.hypotheses.org/>.

5. Lou Burnard, *Qu'est-ce que la Text Encoding Initiative ?*, Marseille, France, 2015, URL : <http://books.openedition.org/oep/1237> (visité le 15/08/2022), p. 3-4.

environnement numérique en pleine expansion⁶. En raison de sa structure modulaire et de sa grande flexibilité, la TEI met à notre disposition un vaste ensemble de balises sémantiques nous permettant de décrire la structure logique du dictionnaire et d'en indexer les entités signifiantes.

Cette flexibilité doit cependant être restreinte dans un schéma documenté afin de répondre aux besoins spécifiques du projet et d'assurer un uniformité et intégrité des données⁷. Ce document associant documentation et schématisation ou adaptation formelle des règles de la TEI est qualifié d'O.D.D⁸. En conséquence, le premier enjeu clé de notre travail a été de proposer un modèle d'encodage et son schéma qui soient suffisamment proches des attentes scientifiques du projet, tout en permettant une automatisation de la pose des balises, une réutilisation et transformation des documents puis la pérennisation sur le long terme des données.

Dans un second temps, face à la taille du corpus (311 pages de notices au format DOCX en avril 2022) et son caractère évolutif (d'autres notices devant s'ajouter pour compléter le dictionnaire), un encodage manuel complet n'était nullement envisageable. C'est pourquoi nous avons dû proposer un script d'automatisation de la pose des balises TEI et de transformation du document de travail initial en une collection de fichier au format XML, conforme aux principes de la TEI et à notre schéma d'encodage. Pour ce faire, nous avons eu recours à un script dans le langage de programmation Python devant relever les enjeux suivant :

- Créer une arborescence XML/TEI respectant le modèle d'encodage.
- Automatiser la pose des balises structurantes, d'après des modèles récurrents grâce à des expressions régulières.
- Permettre un encodage différenciant les sources primaires de la bibliographie scientifique.
- Assurer les liens entre les notices par l'encodage des renvois entre les notices, que ce soit dans les titres ou dans le corps du texte.
- Indexer les entités signifiantes dans le cadre du dictionnaire par le biais de la reconnaissance des entités nommées à partir d'une ontologie spécifique.

6. *Ibid.*, p.5.

7. *Ibid.*, Voir chapitre « Personnaliser la TEI ».

8. Acronyme emprunté aux *Guidelines* de la TEI, signifiant *One Document Does it all*, impliquant que ce document présente d'une part en prose les choix d'encodage, et d'autre part la schéma XML/TEI en lui-même.

Ce dernier enjeu fut indéniablement l'un des plus complexe à traiter, dans la mesure où le dictionnaire comprend des toponymes, des anthroponymes et des ergonymes⁹ qui ne peuvent être différenciées simplement par des modèles récurrents distincts¹⁰. En conséquence, nous avons dû compléter notre chaîne de traitement des données par l'insertion d'une phase de reconnaissance des entités nommées afin de satisfaire ce besoin énoncé dans les lignes directrices du projet.

Dans un dernier temps, nous avons proposé un prototype d'application web permettant de pré-visualiser la forme finale du projet et le réseau complexe de renvois entre les notices, dessinant des approches thématiques et géographiques à différents niveaux. L'enjeu sous-jacent étant donc celui de la mise à disposition et présentation des données transformées.

En conséquence, notre travail s'est structuré autour des enjeux d'interopérabilité et de pérennisation des données nativement numérique. Dès lors, il nous faut nous questionner sur les spécificités de ces données produites dans le cadre du dictionnaire. Effectivement, en quoi le caractère historique plus que linguistique du *Dictionnaire de la Ferme générale*, à la croisée des besoins des chercheurs et des impératifs de la science ouverte, implique une modélisation, un encodage et un traitement de la donnée singulier ? De même, dans quelle mesure la mise à disposition des données dans le cadre d'une application web implique une adaptation du processus de traitement aux diverses échelles temporelles du projet ?

Afin de mieux aborder ce questionnement qui fut central dans notre travail, nous proposons de développer trois axes clés présentant et mettant en contexte le travail accompli durant notre stage. Tout d'abord, il nous faut comprendre le contexte épistémologique, historiographique et institutionnel dans lequel s'insère le *Dictionnaire de la Ferme générale* afin de mieux saisir les enjeux spécifiques de l'encodage des données. Dans un second temps, il est nécessaire de présenter en détail les aspects techniques de la schématisation, puis de la chaîne de traitement et d'encodage automatique des données, afin de mieux saisir dans une dernière partie la mise en œuvre d'une ouverture des données par le biais d'une application web.

9. Groupe d'entités nommées comprenant les institutions, organisations et juridictions.

10. Nous employons par la suite l'expression anglo-saxonne consacrée *patterns* pour qualifier ces modèles récurrents d'entités nommées ou d'expressions.

Première partie

Le *Dictionnaire de la Ferme générale* et l'ANR FermeGé, un *workflow* à diverses échelles

Chapitre 1

La Ferme générale comme objet d'étude, de l'ANR au *workflow* du stage

Dans leur adresse à l'Assemblée nationale en date du 9 juillet 1791, les « employés » de Paris de la Ferme générale déclarent : « les préposés à la perception des droits nationaux sont réellement des fonctionnaires publics¹ ». À bien des égards, cette évocation du fonctionnariat dont relèveraient les commis et agents de la Ferme générale cristallise nombre d'enjeux historiographiques. Effectivement, cette institution singulière, en charge de la perception des aides entre 1640 et 1794 tient un rôle particulier dans le paysage des institutions d'Ancien Régime d'une part, et dans l'historiographie d'autre part. Dès lors, nous pouvons nous questionner sur les raisons scientifiques motivant une étude approfondie de la Ferme générale. De même, dans quelle mesure le stage dont est issu ce présent mémoire s'insère dans un programme d'étude à différentes échelles et sur différentes temporalités ?

Si son fonctionnement administratif et le statut de ses commis a été étudié en détail, son emprise concrète sur les territoires et les populations reste relativement difficile à appréhender. Face à au morcellement des connaissances entourant cette institution pourtant centrale, l'ANR FermeGé entend répondre à ce manque historiographique. L'objectif est donc de fournir une analyse relevant de l'« histoire totale », se voulant la plus exhaustive possible. Afin de mieux comprendre le déroulé de ce vaste programme de recherche et la manière dont notre travail s'est inséré dans un *workflow* pluriannuel, il nous semble nécessaire de dresser un rapide portrait historique et historiographique de notre objet de travail. Cette approche historiographique nous invite à mettre en évidence les nouvelles approches thématiques proposées par les directeurs scientifiques du projet. Dès lors, nous pourrons replacer l'ANR FermeGé dans son contexte institutionnel, et mettre en évidence

1. Vida Azimi, *Un modèle administratif de l'Ancien régime : les commis de la Ferme générale et de la Régie générale des aides*, Paris, France, 1987, p.5.

le rôle de la Maison Européenne des Sciences de l'Homme de la Société (Lille) comme structure d'accueil du projet, nous permettant de contextualiser notre travail.

1.1 La Ferme générale comme objet d'étude historique et historiographique

1.1.1 La Ferme générale : genèse, développement, remise en question

D'un point de vue historique, l'histoire de la Ferme générale semble indissociable de celle de la naissance et du renforcement de l'Etat moderne, administratif et financier. En effet, la Ferme générale en tant qu'administration trouve son origine dans le chevauchement et la multitude de fermes particulières au XVI^e siècle. L'historienne Vida Azimi met en évidence la date de 1584 comme première mention de la « ferme générale » lors de l'adjudication à un seul fermier du bail des cinq grosses fermes². Cette réunion des fermes alors éparses ne semble que temporaire puisqu'il est fait mention sous Sully, alors surintendant des finances depuis 1598 de quatre fermes générales distinctes : le rassemblement initial des cinq grosses fermes, la ferme des gabelles de France, de Languedoc et la ferme des aides. C'est indéniablement l'ordonnance sur les fermes de Colbert, datée de 1681, qui marque le début de la structuration de la Ferme générale comme institution cohérente, rationnelle et efficace. Effectivement, avec le bail de Carlier les quatre fermes se voient rassemblées sous l'égide du pouvoir royal. Cette unification des fermes semble être parachevée en 1726 avec le rassemblement de la Ferme générale et celle des droits et du domaine du roi.

Au cours du XVIII^e siècle, la Ferme générale apparaît comme un vecteur de cristallisation des aspirations à une monarchie administrative, mais aussi de sa contestation. En effet, si elle est progressivement érigée en modèle d'administration centralisateur, la Ferme fait l'objet de tentative de remplacement. Sous le système Law (1716-1720), elle est adjugée sur les fonds de la Compagnie des Indes. Néanmoins, l'éclatement de la bulle financière en 1720, qui entraîne dans son sillage l'économiste et son système boursier, implique un retour au modèle de bail pour la Ferme générale. Cette première remise en question de la place de l'institution n'est qu'une première expérience, car elle subit à nouveau une tentative de démembrément sous Turgot, contrôleur général des finances entre 1774 et 1776. Effectivement, cette grande réforme avait pour finalité d'améliorer l'efficacité des agents de la Ferme générale et de réduire les abus financiers. Sous son impulsion, les baux sont résiliés et certains fermages sont détachés de la Ferme générale, tels que les poudres et le

2. *Ibid.*, p.6.

salpêtre ou l'administration des messageries. La chute du ministre n'entraîne cependant qu'un court répit pour l'institution qui se voit démembrée en 1780 avec la création de la Régie générale. Cette nouvelle institution, néanmoins proche dans son fonctionnement de la Ferme générale, récupère l'affermage des aides. La Ferme générale est alors limitée à la perception des gabelles, traites et taxes sur le tabac³. En dépit de son efficacité globale et des supplications de ses agents, la Révolution sonne le glas de la Ferme générale.

1.1.2 La Ferme générale : un « modèle administratif » d'Ancien Régime ?

Un monde de papier, un corps de règles, un instrument de pouvoir, un étalage de bureaux, l'administration est aussi et avant tout une affaire d'homme, de ces bureaucrates⁴.

C'est en ces termes que l'historienne Vida Azimi décrit la Ferme générale dans son étude consacrée au fonctionnement administratif et à ses employés. Il convient donc de mettre en exergue dans un premier temps la hiérarchie administrative de la Ferme générale, incarnée par ses agents, afin de mieux saisir l'emprise territoriale qu'elle exerce. Effectivement, l'organisation hiérarchique de la Ferme générale peut être schématisée de la façon suivante :

- Les **Fermiers généraux** : à la tête de l'institution, ils ont un rôle de contrôle global et d'organisation du travail au sein de la Ferme. Ils doivent être titulaires d'un brevet de Fermier général et coopté par leurs collègues. L'historien Yves Durand, dans son ouvrage consacré au corps des fermiers généraux met en évidence leur extraction sociale relevant de la moyenne voire haute bourgeoisie. La charge de Fermier général étant effectivement un moyen d'accéder à la noblesse⁵.
- Les « **cadres supérieurs** » : catégorie regroupant les directeurs, contrôleurs généraux, inspecteurs et receveur généraux. Leurs domaines d'activités sont avant tout la direction effective et surveillance du bon fonctionnement de l'administration. Ces acteurs administratifs essentiels agissent soit au sein de la Direction générale des fermes à Paris, où ils assurent notamment la transmission et communication des ordres, soit dans une direction provinciale. A une généralité correspond globalement une direction provinciale de la Ferme générale. Dans ce cas, les directeurs provinciaux mettent en application les décisions de la compagnie, assurent la perceptions des droits et mènent des inspections dans les arrondissement ainsi qu'une lutte permanente contre la fraude et contrebande.

3. *Ibid.*, p.8.

4. *Ibid.*, p.6.

5. Yves Durand, *Les Fermiers Généraux au XVIIIe siècle*, Paris, PUF, 1971, p.104

- Les « **cadres principaux** » : contrôleurs, receveurs ou commis supérieurs sont sous les ordres des directeurs provinciaux, ils accomplissent des tâches d'applications, tel que le contrôle particulier sur un bureau des aides. Parmi eux, les receveurs particuliers, pouvant être ambulants, encadrent la perception de la gabelle et des aides sur les huiles, savons, traites et marques de fer.
- Les « **cadres d'exécution** » : commis subalternes et employés des brigades, ils sont en charge de la rédaction des écritures, de la perception manuelle des droits ou de la vente du sel. Dans cette catégorie, les employés des brigades tiennent un rôle singulier puisqu'ils constituent une forme d'armée privée de la Ferme générale, en charge de la lutte contre la fraude et la contrebande armée.

Cette organisation hiérarchisée, rationalisée et contrôlée érige la Ferme générale en un modèle administratif d'Ancien Régime. Cet aspect modélisant de la Ferme est remarquable à travers la volonté de s'assurer du contrôle sur ses agents jusqu'au plus bas niveau de l'échelle administrative. En effet, de manière systématique, les commis et employés de l'administration sont « déracinés » de leur territoire de naissance dans le but d'éviter toute compromission dans des affaires de corruption en faveur d'un proche. D'autre part, un certain nombre d'instruments d'uniformisation sont mis en place en interne, comme par exemple le développement de guide des employés et d'ordre de travail spécifiquement adressés aux commis afin de circonscrire leur périmètre d'action⁶.

1.1.3 La Ferme générale : objet de contestations multiples

Tout comme elle s'affirme en tant que modèle administratif, la Ferme générale concentre les contestations à diverses échelles. Le développement de la Ferme générale et de la fiscalité moderne s'accompagne d'une littérature contestant le bien fondé de la notion d'impôt indirect. Au niveau politique, cette forme de taxation est vue comme une contrainte imposée aux corps intermédiaires par les financiers, dévoyant ainsi la bonne tenue de la fiscalité royale. Cette contestation intellectuelle s'observe dans l'ouvrage de Mirabeau (père), *La Théorie de l'Impôt* dans lequel il prône, dans la continuité des idées physiocratiques naissantes, l'établissement d'un impôt unique et juste en lieu et place des taxes indirectes. La figure du financier à la tête de la Ferme générale est donc vivement critiquée sous sa plume :

Il ne faut que supprimer ce mot odieux : « financier »⁷

6. Par exemple, la rédaction sur ordre du fermier La Motte des instructions précises aux employés pour leur apprendre à correctement verbaliser *Ibid.*, p.61.

7. Mirabeau, *Théorie de l'Impôt*, Paris, 1761. Cité par Vida Azimi, *Ibid.*, p. 74

Au niveau administratif et judiciaire, les méthodes employées par la Ferme générale et ses brigades font l'objet de vives critiques. Il semble que les relations entre les commis, personnages « déracinés » et donc extérieurs à la communauté soient tendues et conflictuelles. En effet, les commis ont régulièrement recours à la dénonciation pour endiguer la contrebande endémique. De plus, s'ajoute le droit de visite des commis et les lourdes peines infligées aux fraudeurs. L'ordonnance de Colbert de 1680 prévoit en effet des peines allant du coup de fouet à la galère à vie. En somme, le commis et le brigadier, têtes de pont de la Ferme générale, revêtent la figure de l'étranger dans les communautés locales qui apporte avec eux les taxes indirectes, contraires aux coutumes. Les termes de « gabeleurs » ou « gabelous » qui leurs sont associés révèlent indéniablement cette hostilité latente.

Dans une dernière analyse, cette contestation de la Ferme générale se traduit par le développement de la contrebande armée qui connaît des piques d'activités, notamment dans les années 1750⁸. Rappelons brièvement que la contrebande est considérée comme un cas royal car crime contre les lois de l'Etat. Sa répression judiciaire relève donc de la justice retenue du roi qui s'exprime par le biais des priviléges judiciaires de la Ferme générale. Par exemple, l'ordonnance des fermes de 1687 accorde un arsenal de droits de confiscation à l'encontre des fraudeurs et au profit de la Ferme générale. A bien des égards, le XVIII^e siècle semble marqué par une forme de professionnalisation de la contrebande armée, comme en témoigne les actions de grande envergure menées par le célèbre Mandrin en 1756. La lutte contre la contrebande armée est relativement difficile pour la Ferme générale qui est d'une part contrainte de militariser ses brigades, et d'autre part de créer des zones tampons aux frontières des généralités, qualifiées de « lieues limitrophes » où l'implantation des lieux de vente est interdite. En tout état de cause, la Ferme générale endosse un aspect bicéphale, à la fois modèle d'administration financière par son efficacité, et repoussoir car facteur de développement des taxes indirects honnies par les communautés locales. C'est toute cette ambiguïté que l'ANR FermeGé entend saisir par l'apport d'une analyse historiographique reposant sur des couples thématiques.

8. Jean Nicolas, *La rébellion française : mouvements populaires et conscience sociale*, Paris, 2002.

1.2 La Ferme générale comme objet d'étude dans le cadre de l'ANR : le projet FermeGé

1.2.1 L'ANR FermeGé : contexte, objectifs et fonctionnement

Le projet ANR FermeGé s'est donc constitué dans le cadre de ce constat d'une carence historiographique sur ce sujet. Effectivement, comme nous l'avons montré, le fonctionnement administratif et judiciaire de la Ferme générale est relativement bien connu des historiens. Néanmoins, en raison du peu de synthèse sur cet objet historique complexe, ou de leur ancienneté, il semblait nécessaire de proposer une étude englobante, faisant appel à l'histoire politique et des institutions, tout comme aux apports plus récents de l'histoire judiciaire et culturelle. Comme le rappelle le résumé de soumission de l'ANR⁹, les objectifs scientifiques de cette étude de grande ampleur sont les suivants :

- étudier l'impact de cette organisation, discriminante mais rationnelle, sur les territoires et les communautés.
- comprendre l'insertion de la Ferme générale dans la culture du privilège d'une part et dans l'aspiration à l'efficacité administrative d'autre part.
- mettre en lumière les rapports complexes entre les sociétés, les impôts indirects et les territoires, entre coercition et compromis.
- proposer une étude à diverses échelles, notamment à l'échelle locale par le dépouillement d'archives inédites au sein des centres d'archives départementaux notamment.
- mettre à disposition et restituer ces connaissances et nouvelles approches à la communauté scientifique en s'inscrivant dans une approche de science ouverte.

Lancé en septembre 2021 et financé pour 48 mois, le projet se structure autour de quatre axes principaux :

- « **Axe 1 : Dictionnaire numérique de la Ferme générale, objet d'histoire totale** ». Sous la direction de Marie-Laure Legay et Thomas Boullu, l'axe 1 repose sur la réalisation d'un dictionnaire numérique de Ferme générale au fil de l'eau. L'objectif est ici triple, puisqu'il s'agit de réunir en une seule oeuvre les monographies locales et études thématiques concernant la Ferme générale ; de renouveler l'historiographie en interrogeant des thématiques peu explorées jus-

9. <https://anr.fr/Projet-ANR-21-CE41-0019>

qu’alors, telles que l’impact du contrôle des actes sur les administrés, la défense des contrebandiers contre les commis par les inscriptions en faux et la dialectique sous-jacente du rapport institution/individu à double-sens ; de proposer une exhaustivité géographique par la diversité des territoires traités. Notre travail s’est donc inséré dans cet axe à la confluence des enjeux techniques et des perspectives scientifiques.

- « **Axe 2 : L’atlas numérique de la Ferme générale** ». Coordonnée par Benjamin Furst¹⁰, entend développer une atlas numérique de l’ancrage territorial et géographique de la Ferme générale, tant sur les provinces que les bassins fluviaux et les frontières, considérées comme interfaces de commerce et de contrebande.
- « **Axe 3 : Une histoire exploratoire et transdisciplinaire du binôme notionnel inégalité/rationalité** » Orchestré autour d’un cycle de conférence, de publication d’articles scientifiques et de colloques annuels, l’objectif de cet axe scientifique est de proposer un espace de croisement et de confrontation des approches disciplinaires concernant la Ferme générale.
- « **Axe 4 : Un ancrage dans les principes de la science ouverte** ». Cet axe coordonne les enjeux scientifiques et surtout techniques d’interopérabilité, pérennisation et d’ouverture des données de la recherche associées à ce projet. Cet axe transversal a donc été une des composante importante de notre travail, tant en amont sur la modélisation et l’encodage des données que dans leur traitement et mise à disposition des notices.

1.2.2 La Ferme générale : apports et enjeux historiographiques

Afin de clore cette approche organisationnelle et ce bilan historiographique, il nous faut mettre en évidence les apports thématiques et la méthodologie développée au cours du projet afin de constituer les notices scientifiques du dictionnaire, constituant notre corpus de données. L’axe 1 du projet s’est constitué autour de la rédaction de notice scientifique au sein d’un document numérique au format DOCX (*Word*). Ces notices ont d’abord été publié sur le carnet Hypothèses.org du projet que l’on peut consulter à l’adresse suivante : <https://dicofg.hypotheses.org/>. Cependant, dans le cadre de l’ANR et de la structuration de l’axe 2, le choix a été fait d’arrêter cette publication au fil de l’eau afin de proposer un dictionnaire nativement numérique qui puisse répondre à l’ensemble des enjeux techniques, scientifiques et de gestion des données. Au moment du début de notre travail, nous disposions donc des notices sur le carnet de recherche en ligne et d’un

10. Ingénieur de recherche cartographe, UR 3436 CRESAT, Université de Haute-Alsace

document Word constitué de 311 pages ou 299 notices. Cela nous permet donc d'avoir une vision relativement complète des thématiques abordées dans le corpus de notice au commencement du stage. En avril 2022 le corpus initial se structure autour de trois types de notices répondant aux enjeux scientifiques du projet¹¹ :

- les **notices « géographiques »** (107), traitant des provinces, villes ou toponymes et permettant de comprendre l'emprise territoriale de la Ferme générale.
- les **notices concernant les taxes et droits** prélevés par la Ferme générale (80), dans une perspective d'histoire financière.
- les **notices traitant du fonctionnement de l'institution** et de son administration (69), relevant principalement de l'histoire du droit et des institutions.
- les **notices attachées à décrire les rôles des agents et du personnel** de la Ferme, la majorité étant des notices biographiques (34).
- les **notices traitant des différentes modalités de contestations** (7), incluant les actions de refus de versement des impôts, de contrebande ou de révolte ouverte¹².

Soulignons d'ores et déjà que la particularité de ce corpus, et par extension des données sur lesquelles nous avons travaillées, sont évolutives et ont pour vocation d'être complétées. Cette spécificité implique un certains nombre de choix dans le traitement de la donnée et dans la modélisation, ainsi qu'un travail par itérations, que nous détaillerons par la suite. Cependant, en avril 2022, le dictionnaire en constitution devait encore intégrer certaines notices et approches. Par exemple, madame Marie-Laure Legay, coordinatrice du projet et professeure des universités, nous avait effectivement indiqué que le Sud du royaume de France était encore peu traité et des notices telles que Bordeaux ou Lyon devraient alors être rapidement intégrées.

A présent que nous avons établi le cadre historiographique et organisationnel, il convient de présenter brièvement le contexte institutionnel et, un tant soit peu, matériel dans lequel s'insère notre travail.

11. Soulignons le fait que nous proposons une typologie *a posteriori* à partir de nos observations du corpus. Une typologie plus approfondie doit être développée par l'équipe scientifique dans la suite du projet

12. Nous ne considérons pas comme une catégorie à part entière la notice relevant de l'histoire du clergé (« Clergé ») et de l'histoire militaire (« soldats »)

1.3 Le projet FermeGé et sa structure d'accueil : la MESHS comme noeud d'un réseau infra-structurel

Afin de proposer une vision complète du contexte initial dans lequel notre travail s'est inséré, il nous faut mettre en exergue les spécificités des infrastructures mobilisées par le projet à ses diverses échelles.

1.3.1 La MESHS : infrastructure d'accueil et d'appui à la recherche

Notre stage s'est donc déroulé au sein de la Maison Européenne des Sciences de l'Homme et de la Société (MESHS), située à Lille (Nord), au 2 rue des Canonniers¹³. La MESHS est donc une « Unité d'Appui à la Recherche » (UAR 3185), depuis janvier 2022, placée sous la tutelle du CNRS et des établissements d'enseignements supérieur et de recherche de la région des Hauts-de-France : l'Université de Lille (UdL), d'Artois, du Littoral et de la Côte d'Opale (ULCO), l'Université Polytechnique de Valenciennes, l'Université de Picardie Jules Verne (UPJV) et l'Université Catholique de Lille (La Catho). Elle propose un ensemble de moyens d'actions et de services de soutien à la recherche en Sciences Humaines et Sociales (SHS) dans la région. D'un point de vue historique, la MESHS a connu une genèse particulière puisqu'elle est née de la fusion entre l'Institut Fédératif de Recherche - Économie et Sociétés Industrielles (IFRESI) et l'Institut international Erasme en 2008. Cette origine duale oriente donc son action « européenne » par les liens avec les réseaux de recherche du Nord de l'Europe, ou plus concrètement par l'accueil de chercheurs internationaux. En tout état de cause, la MESHS se trouve au cœur d'un dense réseau de laboratoires régionaux puisqu'elle fédère 40 unités de recherche en SHS, soit un potentiel de 2000 chercheurs et ingénieurs¹⁴.

La carte suivante met en évidence l'implantation de la MESHS dans ce réseau régional des laboratoires :

13. Une présentation plus complète de l'organigramme de fonctionnement de l'unité se trouve à l'adresse suivante : <https://www.meshs.fr/page/presentation>

14. Nous remercions l'équipe chargée de la communication de la MESHS pour nous avoir transmis les précieux supports du MESHS Tour de mai 2022, nous permettant d'approfondir notre propos et de l'illustrer dans cette partie

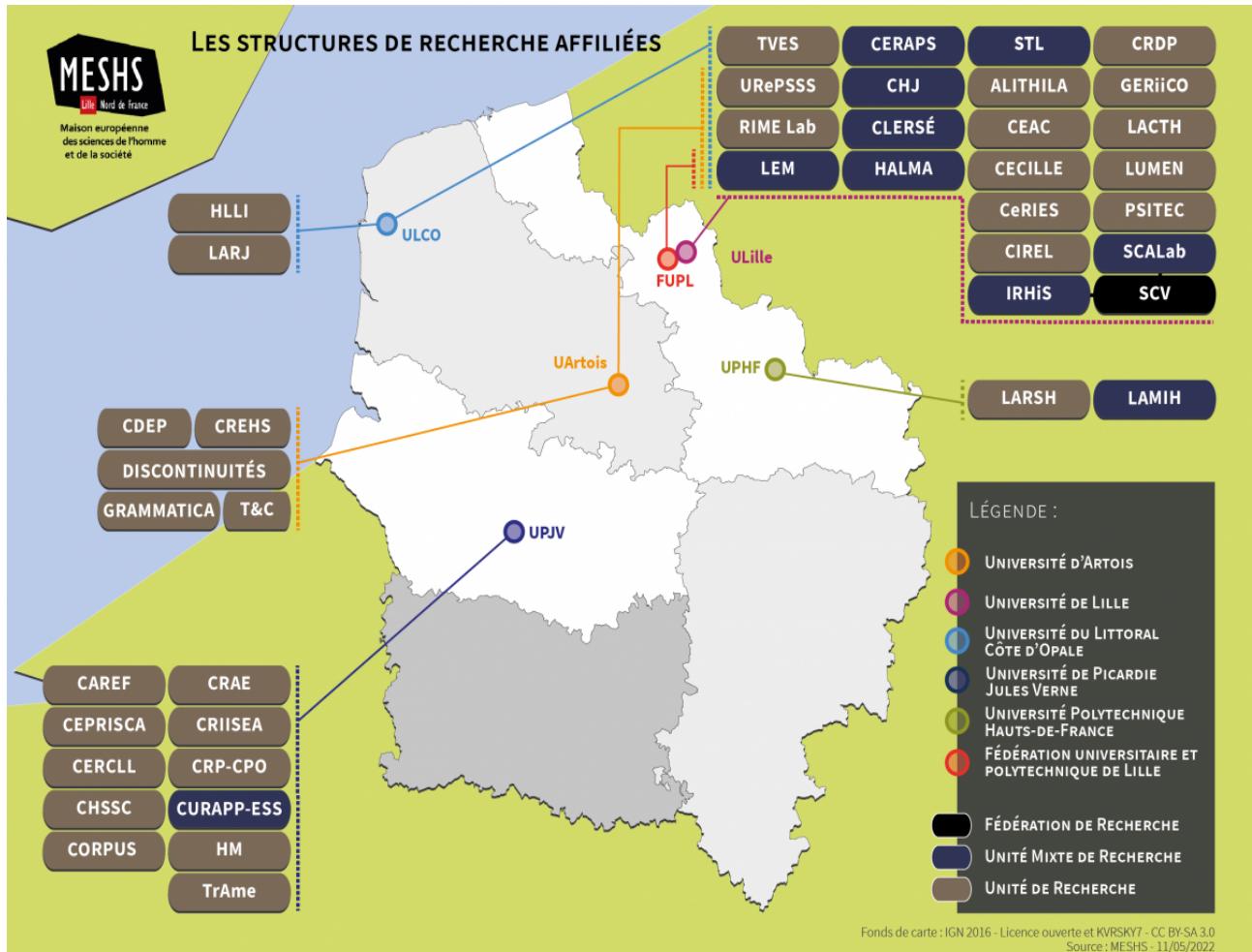


FIGURE 1.1 – Carte des laboratoires affiliés à la MESHS, 2022.

De même, à l'échelle nationale, l'unité s'insère pleinement dans le Réseau National des Maisons des Sciences de l'Homme (RnMSH). Ce « groupe d'intérêt scientifique » a pour mission principale de mettre en commun les savoirs et compétences dans une perspective d'interdisciplinarité. Ce réseau regroupe donc les 22 Maisons des Sciences de l'Homme (MSH) françaises et s'affirme comme un relais essentiel entre les besoins locaux de la MESHS en terme de diffusion des connaissances ou de mutualisation des infrastructures, et les Très Grandes Infrastructures de Recherche (TGIR) telles qu'HumaNum. Cette TGIR, requalifiée en 2022 d'Infrastructure de Recherche * (lire « étoile » ou IR*) organise par l'intermédiaire de son consortium la communauté de recherche en humanités numériques. Cette action s'affirme aussi par le développement plateformes de gestion des données dans le cadre des principes FAIR (Facile à trouver, Accessible, Interopérable et Réutilisable), de mise à disposition d'outils de traitement des données et de serveurs. En conséquence, d'un point de vue institutionnel, la MESHS est une structure d'appui essentielle pour les projets de recherche en SHS disposant d'enjeux de traitement et gestions des données. Il nous faut à présenter la manière dont le projet ANR FermeGé s'est inséré dans ce rouage.

1.3.2 Le réseau d'institutions mobilisées pour le projet

L'ANR FermeGé est donc accueillie pour ses activités et missions d'ingénierie au sein du pôle Humanités numériques, outils, méthodes et analyse de données (HNOMAD) de la MESHS. L'équipe attachée au projet est composée de Victoria Le Fournier (directrice de stage) et Florence Perret, ingénieries chargées du traitement des données scientifiques. L'équipe est en étroite collaboration avec Marie-Laure Legay, coordinatrice du projet, et Thomas Boullu, directeur de l'axe 1 en charge de la réalisation du dictionnaire numérique. Les coordinateurs et directeurs ont donc un rôle de liaison entre les activités quotidiennes au sein du pôle Humanités Numériques de la MESHS et l'ensemble du réseau de partenaires institutionnels et universitaires associé à l'ANR. Effectivement, le projet associe notamment le Centre d'Histoire et d'Anthropologie du Droit (CHAD), le Centre de Droit Privé et de Sciences Criminelles d'Amiens (CEPRISCA), le Centre de Recherches sur les Economies, les Sociétés, les Arts et les Techniques (CRESAT) et l'Institut de Recherche Historique du Septentrion (IRHiS). C'est donc à nouveau une articulation du projet à diverses échelles institutionnelles que nous pouvons mettre en évidence.

Cette insertion dans un réseau national doit être complétée par la mention de la collaboration avec l'IR* HumaNum qui met à disposition du projet un ensemble d'outils concernant le « cycle de vie » des données du dictionnaire numérique. Effectivement, que ce soit par le biais des outils de « versionnage » des documents textes servant de corpus initial ou des plateformes de dépôts et archivage des données, les infrastructures d'HumaNum sont sollicitées tout au long du projet. Le schéma suivant mets en évidence la manière dont nous avons utilisé ces différents outils au cours de notre stage :

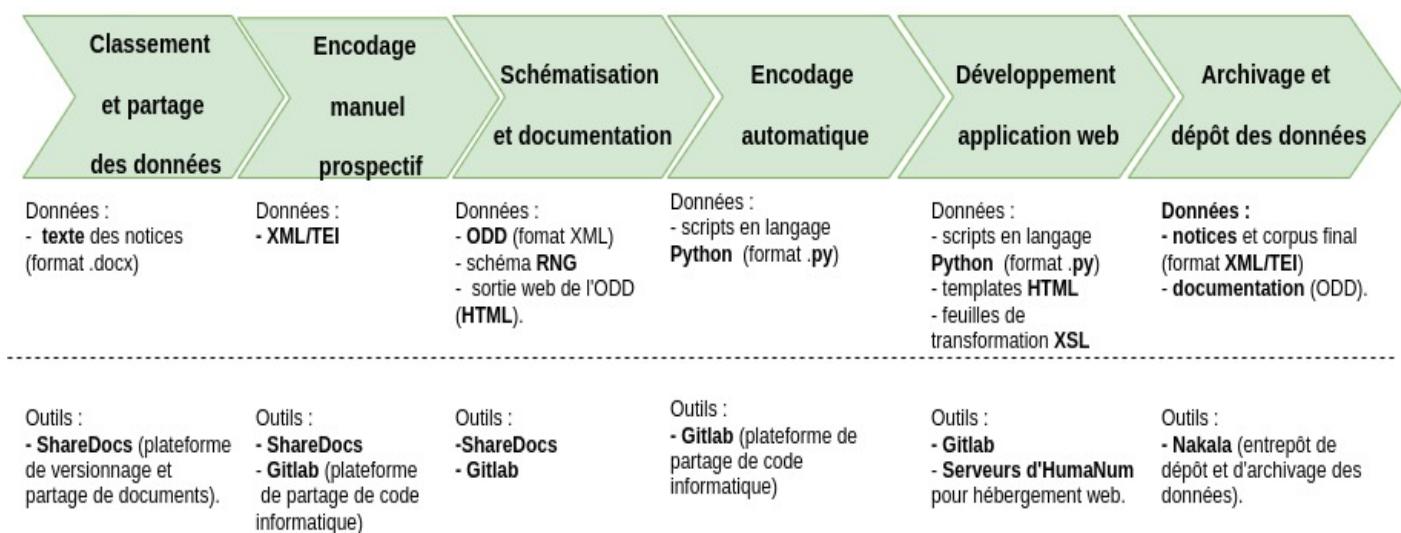


FIGURE 1.2 – Diagramme des outils d'HumaNum utilisés dans le *workflow* du projet.

En conséquence, le projet ANR FermeGé s'insère indéniablement dans un réseau dense d'institutions et laboratoires de recherche, que ce soit au niveau de la constitution du corpus de notices, du traitement des données ou de leur mise à disposition des données

de la communauté scientifique. Il nous faut donc dans un dernier temps indiquer la manière dont notre travail s'est inséré dans la temporalité du projet.

1.3.3 Le stage dans son contexte temporel et institutionnel

Notre travail s'intègre donc dans la temporalité d'un projet de recherche de cinq années. Un calendrier prévisionnel a été établi au début du projet, donnant à voir la manière dont les différents axes interagissent. Notre stage a donc pris place à l'intersection des tâches 1-2 et 1-3, à savoir la coordination et l'encodage des notices.

A l'échelle du stage en lui-même, 5 tâches que nous avons traités consécutivement pour les deux premières et par itérations pour les autres nous ont été confiées :

- Tache 1 : Gestion des fichiers XML
- Tache 2 : Etat de l'art sur la question d'une édition de dictionnaire avec la TEI.
- Tache 3 : Élaboration du modèle d'encodage en coopération avec M.-L. Legay et Th. Boullu.
- Tache 4 : Automatisation de la pose de balises dans les notices.
- Tache 5 : Intégration du modèle de données sur le CMS choisi.

En conclusion, il nous faut souligner le fait que notre travail s'inscrit dans un projet ANR mobilisant des acteurs à plusieurs niveaux institutionnels. Ainsi, la question de la temporalité des diverses tâches au cœur du projet est cruciale. C'est pourquoi nous proposons d'analyser et de présenter dans les chapitres suivants les tâches que nous avons traitées. En premier lieu, l'état de l'art sur l'encodage des dictionnaires en XML/TEI et l'encodage prospectif des notices ont été des étapes indispensables pour pouvoir développer notre schéma d'encodage, puis un processus d'encodage automatique. En somme, nous présentons d'abord les enjeux d'encodage et de traitement de la donnée auxquels nous avons été confrontés dans les chapitres 2 et 3 pour ensuite mettre en perspective les solutions techniques mises en oeuvre dans les chapitres suivants.

Chapitre 2

Encoder un « dictionnaire nativement numérique » en TEI : enjeux, perspectives et limites

In other cases [...] definitions bear testimony of the evolution of society [...]¹.

C'est sur cette affirmation que le groupe d'ingénieurs, de linguistes et d'historiens composé d'Hervé Bohbot, Francesca Frontini, Giancarlo Luxardo, Mohamed Khemakhem et Laurent Romary conclut la présentation du projet Nénufar, visant à proposer un encodage en XML/TEI et une édition numérique du *Petit Larousse Illustré* (jusqu'en 1948). Comprendons donc que les dictionnaires apparaissent non seulement comme un corpus de notices recensant les connaissances, mais aussi comme un artefact ou une photographie d'une société et de sa culture. Si cet aspect semble indéniable pour les dictionnaires historiques, par la suite encodés et numérisés, la situation est-elle comparable pour un dictionnaire nativement numérique, produit par une communauté de chercheurs ? Ainsi, ce présent état de l'art se propose de fournir une analyse des enjeux épistémologiques et historiographiques concernant l'encodage des dictionnaires en XML/TEI, afin de mettre en perspective les particularités et les choix appliqués au *Dictionnaire de la Ferme Générale*. Effectivement, dans quelle mesure l'encodage en XML/TEI des dictionnaires permet de lier sémantisme et mise à disposition des données ? En quoi le caractère nativement numérique du dictionnaire de la Ferme Générale lui permet de s'inscrire dans un réseau dense de projets, à la croisée de l'encodage des dictionnaires et des bases de données textuelles ?

Dans un premier temps, il nous faut comprendre les enjeux liés à l'encodage des dic-

1. Traduction : « En d'autres cas, les définitions traduisent les évolutions d'une société » Hervé Bohbot, Francesca Frontini, Giancarlo Luxardo, Mohamed Khemakhem et Laurent Romary, « Presenting the Nénufar Project : a Diachronic Digital Edition of the Petit Larousse Illustré », dans *GLOBALEX 2018 - Globalex workshop at LREC2018*, Miyazaki, Japan, 2018, p. 1-6, URL : <https://hal.archives-ouvertes.fr/hal-01728328> (visité le 12/04/2022), p.4.

tionnaires et la manière dont le *Dictionnaire de la Ferme Générale* s'inscrit dans cette constellation de travaux. Dès lors, dans un second temps, nous pouvons mettre en perspective les choix d'encodage et de schématisation des données du Dictionnaire pour analyser ses spécificités et enjeux, notamment en lien avec l'indexation et l'encodage des entités nommées.

Afin de mettre en relation ce chapitre avec le travail que nous avons effectué au cours de ce stage, nous proposons un schéma évolutif, complété au fil des chapitres de ce mémoire, du *workflow* que nous avons mis en place, dont voici la première étape :

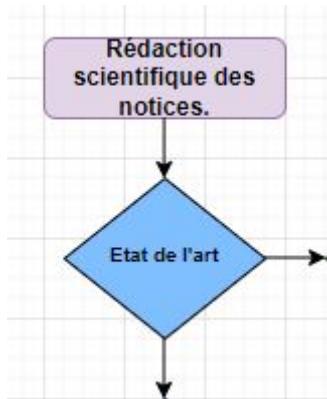


FIGURE 2.1 – *Workflow* : état de l'art de l'encodage des dictionnaires.

2.1 Panorama épistémologique et historiographique de l'encodage des dictionnaires

2.1.1 Structuration de l'information dans un dictionnaire

En premier lieu, un dictionnaire se présente sous la forme d'un document structuré par des entrées, portant *a minima* deux types d'informations encyclopédiques : le « libellé » d'un concept, et une description exhaustive pouvant comporter des informations scientifiques, historiques, linguistiques, ect (la définition)². Dès lors, comme le soulignait Nancy Ide³, fondatrice du module appliqué aux dictionnaires et rédactrice du chapitre associé dans les *guidelines* de la TEI dans leur troisième version (P3) en 1993, les dictionnaires sont à la fois des textes et des bases de données, ce qui implique une profonde difficulté de structuration de l'encodage. Effectivement, le lecteur accède à l'entrée recherchée par une vedette (la « clé ») dans le but de récupérer le contenu informationnel (la

2. M. Khemakhem, Simon Gabay, Béatrice Joyeux-Prunel, L. Romary, Léa Saint-Raymond et Lucie Rondeau Du Noyer, « Information Extraction Workflow for Digitised Entry-based Documents », dans *DARIAH Annual event 2020*, Zagreb / Virtual, Croatia, 2020, URL : <https://hal.archives-ouvertes.fr/hal-02508549> (visité le 12/04/2022), p.2.

3. Nancy Ide, Ab Et et Jean Véronis, « Codage TEI des dictionnaires électroniques », *Cahiers Gutenberg* (, 1^{er} janv. 1996), DOI : 10.5802/cg.197, p.3.

« valeur »). Cette dualité structurelle clé-valeur est renforcée par l'organisation hiérarchique des dictionnaires traditionnels sous forme de niveaux emboîtés. Les caractéristiques et informations du niveau hiérarchique de l'entrée sont « héritées » par les niveau inférieurs ou enfants, tels que la forme prise par le lemme. Les données contenues à un certain niveau de hiérarchisation ont donc une portée, tout comme les variables d'un langage informatique.

2.1.2 Les dictionnaires et les métalangages d'encodage

Ainsi, il n'est pas surprenant que les dictionnaires, ou documents structurés par des entrées (« entry-based documents »), furent au cœur des enjeux d'encodage de l'informatique naissante. Comme le relate les chercheurs en traitement automatique des langues (TAL) Mathieu Mangeot et Chantal Enguehard, dès les années 1980, deux types de dictionnaires électroniques ont vu le jour : d'une part les dictionnaires compilés d'application informatique, et d'autre part les bandes de photo-composition⁴. Ce dernier type de document était l'objet d'ambiguïté d'interprétation car la structure était relativement lâche et nécessitait une *Document Type Definition* (DTD⁵) particulière pour chaque document. L'eXtended Markup Language (XML), développé en 1997 par le W3C, fut rapidement adopté pour l'encodage des dictionnaires électronique car il impliquait un changement de paradigme : les dictionnaires électroniques n'avaient plus uniquement pour seul but d'être imprimés, mais pouvaient dès lors être partagés de manière nativement numérique. Une séparation entre le fond et la forme devenait possible, ce qui autorisa la représentation standardisée en XML des dictionnaires. En conséquence, le développement du module concernant les dictionnaires pour le TEI visait à proposer une standardisation rendant explicite la structure informationnelle implicite des dictionnaires non encodés. En conséquence, les *guidelines* de la TEI dans leur cinquième et actuelle version propose un schéma d'encodage dont voici la version minimale, qui reprend cette dualité entre la forme prise par le concept (la « clé »), et le sens (la « valeur ») :

4. Mathieu Mangeot et Chantal Enguehard, « Des dictionnaires éditoriaux aux représentations XML standardisées », *Nuria Gala et Michael Zock. Ressources Lexicales : contenu, construction, utilisation, évaluation* (, 2013), Publisher : John Benjamins, p. 24, DOI : 10.1075/lis.30.08man.

5. Il s'agit d'un fichier décrivant l'utilisation des balises et autres règles d'encodage pour un document spécifique

```

<entry>
  <form>
    <orth> Exemple </orth>
  </form>
  <sense>
    Définition du concept.
  </sense>
</entry>

```

FIGURE 2.2 – Modèle d’encodage minimal en TEI des notices de dictionnaires

2.1.3 Premières réflexions sur le *Dictionnaire de la Ferme générale*

Ce premier constat épistémologique est applicable au *Dictionnaire de la Ferme Générale*. Effectivement, le document de travail de base est constitué d’un fichier Word où la structure des données est implicitement mise en évidence par la typographie comme le montre la notice suivante :

Ferme générale

« Ferme générale » au singulier est un terme utilisé dans les traités que le roi signait avec un *consortium* de financiers à qui il confiait la « ferme générale » de tel ou tel impôt. Ainsi, le terme « Ferme générale des gabelles de France » ou « Ferme générale des aides » est mentionné dès le règne de Louis XIII. Les réunions successives des fermes du roi ont permis de désigner sous l’acception singulière un ensemble de fermes concernant plusieurs types d’impôts. Ainsi, dès 1680, par le bail signé le 27 juin pour six ans, on emploie le terme de « *La ferme générale* des gabelles de France et des évêchés de Metz, Toul et Verdun, Domaines et salines de Lorraine et du Comté de Bourgogne, aydes de France, entrées de Paris et de Rouen, droits sur le papier et parchemin timbrez, cinq grosses fermes, tabac, estain, Douanes de Lyon et Valance (sic), Convoy et comptable de Bordeaux, Patente de Languedoc, et autres fermes et droits y joins ». Contrairement à ce qu’énonce Jean Clinquart, le singulier s’emploie donc dans le langage avant la formation du bail Carlier de 1726 pour désigner la compagnie financière dont nous traçons l’histoire. En 1686 par exemple, François Rémond, intéressé dans le bail Fauconnet, se défend d’être « complice de divertissement de deniers de *la Ferme générale* dont on dit que Jean Gruslé est coupable ». A chaque bail général correspondait une « ferme générale ». Le bail général se déclinait en multiples baux particuliers de fermes et sous-fermes, « réunies » selon diverses modalités.

[*A Monsieur Belin, conseiller du roi en son Châtelet de Paris, rapporteur du procès. (Requête de François Remond, sieur de Bréviande, l’un des intéressés au bail général des fermes unies sous le nom de Jean Fauconnet, poursuivi comme complice des détournements commis par Jean Gruslé au préjudice de la ferme générale)*, février 1686, 16 p. ; Jean Clinquart, « Ce que nous ignorons des fermes générales », *Histoire institutionnelle, économique et financière : questions de méthode (XVIIe-XVIIIe siècles)*, Vincennes, CHEFF, IGPDE, 2004, p. 91-10]

FIGURE 2.3 – Notice « Ferme générale » dans le document de travail du Dictionnaire.

Afin de comprendre les spécificités de l'encodage des dictionnaires, il nous faut replacer le projet dans un réseau plus vaste de travaux d'encodage semblables au niveau national et européen.

2.2 Les dictionnaires et la TEI : une approche et un encodage linguistique prédominant

2.2.1 De la linguistique au traitement automatique des langues

Soulignons tout d'abord que les dictionnaires encodés en XML/TEI sont le fait de travaux relevant de la linguistique, de l'étymologie et plus récemment du traitement automatique des langues. En effet, certains dictionnaires imprimés ont d'abord fait l'objet d'un encodage en TEI, insistant sur les données et métadonnées lexicales et linguistiques d'une part et l'ajout d'une couche sémantique à la structure typographique initiale d'autre part. Citons à cet égard le cas du *Dictionnaire de Ducange en ligne*, publié et maintenu par Frédéric Glorieux, responsable des publications électroniques à l'École Nationale des Chartes (ENC) entre 2007 et 2011⁶. Ce projet se base donc sur la conversion et l'encodage en XML/TEI du célèbre dictionnaire imprimé du latin médiéval⁷. L'objectif du projet était de fournir un enrichissement des données du dictionnaire pour permettre une exploitation linguistique et lexicographique numérique. Cette perspective explique donc l'encodage systématique des abréviations, des citations et de la bibliographie. De plus, le caractère peu systématique des articles du *Glossarium* a déterminé le choix d'encoder les blocs textuels définitionnels avec la balise <dictStrap>⁸, comme le montre l'article suivant :

6. Le schéma d'encodage du dictionnaire et sa documentation sont maintenus à l'adresse suivante : <http://svn.code.sf.net/p/ducange/code/xml/ducange.html>

7. Charles Du Cange et al., *Glossarium mediae et infimae latinitatis*. Niort : L. Favre, 1883-1887

8. La documentation du projet explique effectivement que les aliénas du *Glossarium* ne doivent pas être interprétés comme un marqueur de sens car peu systématiques. Définition de l'élément <dictStrap> tiré des *Guidelines* de la TEI : « encloses a part of a dictionary entry in which other phrase-level dictionary elements are freely combined » (traduction : "englobe la partie d'une entrée de dictionnaire dans laquelle d'autres éléments de niveau "expression" sont librement associés). Voir (<https://tei-c.org/release/doc/tei-p5-doc/en/html//ref-dictScrap.html>)

The screenshot shows the XML/TEI encoding of the Putagium entry from the Ducange online dictionary. The XML code is on the left, and the corresponding Latin text is on the right. The XML uses various TEI tags like <entry>, <dictScrap>, <form>, <quote>, <gloss>, and <cb> to structure the text. The Latin text discusses Putagium's meaning as a prostitute and its etymology from Putagio.

```

<entry xml:id="PUTAGIUM"
rend="ducange">
<dictScrap xml:id="PUTAGIUM-1"><form rend="b">PUTAGIUM</form>,
Fornicatio, meretricatus, de fœmina dicitur. Regiam Majestatem lib. 2. cap. 49 :
<quote>Quod generaliter dici solet, quod Putagium hæreditatem non adimit,
intelligitur de Putagio matris, quia filius est hæres legitimus, quem nuptiæ
demonstrant.</quote> Vide Bractonum lib. 2. cap. 37. § 6. et Fletam lib. 1.
cap. 15. § 4. Gallis <hi rend="i">Putage</hi>. Charta ann. 1247. in Tabular.
Campaniæ fol. 343 : <quote>La fame, qui dira vilonnie à autre, si come de
Putage, payera 5. sols, ou portera la pierre toute nuë en sa chemise à la
Procession, et celle-là poindra aprés an la nage d'un aquillon, et s'elle disoit
autre villonnie, qui atourt à honte de cors, ele paieroit 3. sols, et li homs ainsin.
</quote> Le Roman <hi rend="i">de Vaccès</hi> MS. : <quote><1>Maint
home a essillié et torné à servage,<1>Et mis par poureté mainte feme au
Putage.</1></quote> [...].</dictScrap>
<dictScrap xml:id="PUTAGIUM-2" rend="carpentier">Glossar. Provinc. Lat.
ex Cod. reg. 7657 : <gloss>Putaneiar, Prov. fornicari.</gloss> Charta <cb
n="577c"/> Phil. III. ann. 1283. in Reg. 61. Chartoph. reg. ch. 389 :
<quote>Dimisimus Majori et communitati villæ de Vernolio emendas et
cognitiones, quas in dicta villa et ejus livres habebamus in casibus
infrascriptis, videlicet.... de Putagio, de maledictis, de pravis denariatis
carnium, piscium, etc.</quote></dictScrap>
<!-- [...] -->
</entry>
```

FIGURE 2.4 – Encodage XML/TEI de la notice PUTAGIUM du Dictionnaire de Ducange en ligne.

2.2.2 Les dictionnaires encodés en TEI : quelques projets d'encodage

S'il s'agit ici de l'encodage d'un document non nativement numérique, ce projet reste néanmoins intéressant pour le *Dictionnaire de la Ferme générale* car il met en exergue la nécessité d'un choix d'encodage permettant de rendre explicite la structure sémantique initiale. D'autres projets d'encodage de dictionnaires imprimés vers la TEI sont remarquables et font de preuve de choix de schématisation notables :

- **Le Dictionnaire Informatique de Chrétien de Troyes (DICT)**, se compose d'un lexique complet de l'écrivain et d'une base textuelle encodée en XML/TEI, permettant d'interroger ses cinq romans. Le projet, dirigé par Pierre Kunstmann (Université d'Ottawa), Hiltrud Gerner (Université de Lorraine) et May Plouzeau (Université de Provence) a été mis en ligne en 2014. L'encodage en TEI propose une version lemmatisée des documents, puisque l'objectif est de fournir une vaste base lexicographique. Les encodeurs ont eu recours massivement à l'utilisation des balises <w> (*word*) associées à un attribut @lemma. L'encodage du DICT n'est pas à proprement adaptable pour le projet FermGé mais reste un exemple

du lien entre une base de donnée et un dictionnaire⁹.

- Le projet Nénufar propose une édition numérique du *Petit Larousse Illustré* incluant les éditions de 1905 à 1947, encodées en XML/TEI. Ce projet a été développé par le laboratoire Praxiling de l’Université Paul Valéry de Montpellier, en partenariat avec l’INRIA. L’encodage de ce projet reprend le principe de partition entre la forme prise par le concept énoncé et son sens. Cela se traduit par l’utilisation, au sein d’une balise englobante `<entry>` des balises `<form>` et `<sense>`, comme le montre l’extrait de la notice encodée ci-dessous. Remarquons que les données étymologiques sont encodées dans la balise `<etym>` dédiée à part entière. Cette structuration, dans sa version minimale, est tout à fait pertinente pour notre dictionnaire puisqu’elle ajoute une couche sémantique au bloc définitionnel¹⁰.

Article NEO de l'édition 1906-001

```

<entry xml:id="néo" n="1906-001_undetermined">
  <form type="lemma">
    <orth>NÉO</orth>
  </form>
  <etym>
    <pc>(</pc>
      du
      <lang expand="grec">gr.</lang>
      <mentioned>neos</mentioned>
      <pc>,</pc>
      <gloss>nouveau</gloss>
      <pc>)</pc>
    </etym>
    <sense>
      <def>préfixe qui a la même signification</def>
      <pc>,</pc>
    </sense>
  </entry>

```

FIGURE 2.5 – Encodage de la notice ”neo” tirée du projet Nénufar

9. Pierre Kunstmann, Hiltrud Gerner et Gilles Souvay, « Le Dictionnaire Électronique de Chrétien de Troyes (DÉCT1) : révision et élargissement », dans *XXVIème Congrès International de Linguistique et de Philologie Romane*, Valencia, Spain, 2010, à paraître, URL : <https://hal.archives-ouvertes.fr/hal-00522494> (visité le 12/04/2022).

10. H. Bohbot, F. Frontini, G. Luxardo, *et al.*, « Presenting the Nénufar Project... ».

- Le projet Basnage¹¹ est développé dans le cadre de l’ANR Basnum et vise à proposer une numérisation et encodage complet en XML/TEI du *Dictionnaire Universel de Furetière* dans sa version de 1701 (poursuivie par Henri Basnage de Beauval). Du point de vue de la schématisation, l’encodage des entrées s’appuie sur une bipartition entre la forme du lemme et le sens de la définition comme le montre l’extrait plus bas. Quant à l’aspect structurel, à chaque lettre correspond un fichier XML/TEI à part entière. Ceux-ci sont insérés au sein d’un corpus général par le biais de la balise englobante <teiCorpus> dans un « fichier mère ». De plus, ce travail d’encodage peut être mis en relation avec le *Dictionnaire de la Ferme Générale* puisqu’il propose un encodage des entités nommées de personnes et une indexation dans l’en-tête TEI (le <teiHeader>) avec une granularité très fine.

```

<entry xml:id="gagnant">
  <form type="lemmaGrp">
    <form type="lemma">
      <orth>GAGNANT</orth>
    </form>
    <pc>, </pc>
    <form type="inflected">
      <orth type="part">ante</orth>
    </form>
    <pc>. </pc>
    <gramGrp>
      <pos>adj.</pos>
    </gramGrp>
  </form>
  <sense>
    <def>Qui gagne au jeu. Les gagnans ont joué contre les perdans, qui se sont raquitez<pc>. </pc>
    </def>
  </sense>
</entry>
  . . .

```

FIGURE 2.6 – Encodage de la notice ”gagnant” tirée du projet Basnage.

En somme, ces différents dictionnaires encodés en XML/TEI proposent une conversion de dictionnaires déjà existants à destination principalement des linguistes. Soulignons brièvement qu’en dépit de la flexibilité et de la modularité de la TEI, les chercheurs en linguistiques ont mis au point un langage spécifique d’encodage des dictionnaires donnant une part importante à l’étymologie, nommé le *Lexical Markup Framework*(LMF)¹². Dans le cadre de l’encodage d’un dictionnaire nativement numérique comme le nôtre où l’aspect linguistique n’est pas le cœur du propos, il nous faut ouvrir notre panorama

11. Présentation du projet Basnage et de l’ANR Basnum par le consortium Cahier : <https://cahier.hypotheses.org/basnage>

12. Le LFM est décrit comme un métalangage séparant les parties lexicales, grammaticales et sémantiques de manière multi-scalaire. M. Mangeot et C. Enguehard, « Des dictionnaires éditoriaux aux représentations XML standardisées »..., p.9.

épistémologique aux travaux d'encodage basés sur d'autres documents structurés autour d'entrées.

2.3 Les *entry based documents* : à la confluence des bases de données et dictionnaires

2.3.1 L'hypothèse *Grobid-dictionnaries*

Le développement de l'apprentissage machine (*machine-learning*) permet notamment le traitement et la segmentation de corpus massifs de documents basés sur des entrées, à l'instar des catalogues de livres et de ventes. La comparaison de ces documents avec le projet FermeGé s'avère intéressante car ceux-ci se présentent sous la forme d'une liste d'entrées (les clés) proposant un bloc informationnel plus ou moins développé (les valeurs)¹³. A cet égard, différents projets d'encodage de catalogues ou corpus similaires ont vu le jour récemment en s'appuyant sur le logiciel de reconnaissance oculaire des caractères et de transcription *Grobid-dictionaries*. Citons par exemple le projet Katabase qui met à disposition une base de données des catalogues de livres, de lettres et d'autographes en circulation sur le marché privé, au XIXe siècle¹⁴. La technologie *Grobid* consiste en l'implémentation d'une librairie d'apprentissage supervisé en Java et de nombreux plug-ins assurant un paramétrage des sorties. *Grobid-dictionaries* fonctionne plus spécifiquement sous la forme d'un modèle en cascade avec une segmentation en trois blocs principaux du document (*Headnote/Footnote*, *Body* et *DictStrap*), puis une segmentation du corps du document permettant la reconnaissance des différentes entrées lexicales, et finalement le *parsing* (analyse syntaxique) des entrées lexicales. Voici un exemple de segmentation automatisée produite par Grobid-dictionaries :

13. L. Rondeau Du Noyer, *Encoder automatiquement des catalogues en XML/TEI, principes, évaluations et application à la revue des autographes de la librairie Charavay*, Mémoire de master TNAH, Paris, Ecole Nationale des Chartes, 2019.

14. Les données du *workflow* sont accessibles ici : https://github.com/katabase/1_OutputData

```

<entry>
  <form type="lemma">
    <orth>yosico ini tnahandi</orth>
  </form>
  <pc></pc>
  <form type="inflected">
    <gramGrp>
      <gram>futuro</gram>
    </gramGrp>
    <orth extent="part">cuico</orth>
  </form>
  ....
</entry>

```

→

```

<entry type="phrase">
  <form type="lemma">
    <orth>yosico ini tnahandi</orth>
  </form>
  <gramGrp>
    <pos>verb</pos>
  </gramGrp>
  <pc></pc>
  <form type="inflected">
    <gramGrp>
      <tns>futuro</tns>
    </gramGrp>
    <orth extent="part">cuico</orth>
  </form>
  ....
</entry>

```

FIGURE 2.7 – Exemple d’encodage automatisé d’une entrée avec Grobid-dictionaries

Grobid-dictionaries s’affirme comme un outil efficace et pertinent de segmentation et de pose de balise automatisée. Néanmoins, il nous semble qu’il ne soit pas utilisable en temps que tel dans le cadre de notre projet. Effectivement, ce modèle s’inscrit dans un paradigme d’analyse étymologique et grammaticale, en raison de ses choix de segmentation, qui ne répond que partiellement aux enjeux de l’encodage du *Dictionnaire de la Ferme générale*.

2.3.2 Des projets à la croisée des dictionnaires et bases de données

En conséquence, il nous semble intéressant d’analyser les projets proposant de même des bases de données au format XML/TEI, puisque notre dictionnaire numérique entend à terme se présenter comme une collection exhaustive de notices reliées entre elles et indexées. Dès lors, nous pouvons analyser les choix techniques appliqués au *Thesaurus Exemplorum Medii Aevi* développé par le groupe ThEMA (EHESS) depuis 2004¹⁵. Ce dictionnaire numérique se présente sous forme d’une base de données collaborative mettant à disposition un corpus d’*exempla* médiévaux. Dans un premier temps développé comme une base de données relationnelle sous MySQL¹⁶ et d’une application web en PHP¹⁷, une profonde refonte récente a permis l’encodage complet de ce thésaurus au format XML/TEI en 2019. D’un point de vue structurel le corpus est stocké dans une balise

15. Jean-Paul Rehr et Marie-Anne Polo de Beaulieu, « Thesaurus Exemplorum Medii Aevi : une base de données collaborative sur les exempla médiévaux », *Humanités numériques*–4 (1^{er} déc. 2021), DOI : 10.4000/revuehn.2630, p.7.

16. Système de gestion de base de données relationnelles.

17. Langage de programmation orienté objet utilisé pour le développement d’application web associée à une base de données.

<teiCorpus> et chaque *exemplum* est encodé dans un fichier TEI spécifique à part entière. Comme une grande partie du contenu produit par le groupe ThEMA consiste en des métadonnées (c'est-à-dire les descriptions des *exempla*), la majeure partie des métadonnées se trouve dans l'élément <teiHeader> de chaque document, en particulier l'élément <sourceDesc> pour la description du document et sa bibliographie, ou l'élément <profileDesc> pour les mots-clés. Ainsi, l'encodage en XML/TEI est non seulement un moyen de structurer et d'indexer les *exemplum* et les entités qui les composent, mais aussi un choix dans l'optique de préserver et archiver les données à moyen et long terme. Effectivement, la transformation de la base de données ThEMA en XML-TEI (par opposition à un modèle de données personnalisé) a fait partie d'un plan visant à assurer la longévité et la réutilisation des données du projet. Huma-Num fournit à cet égard l'infrastructure technique sur laquelle repose la base de données de ThEMA (dans eXist-db¹⁸), puis stocke et met les données à la disposition du grand public. Ainsi, ce choix de transition vers un encodage en XML/TEI permet de mettre en évidence la capacité de préservation des données sur le long terme que nous devons appliquer au *Dictionnaire de la Ferme générale*.

Notons néanmoins que certains projets optent pour un modèle de données sensiblement différents dans cette même perspective de pérennisation et partage. Pour preuve, le *Dictionnaire des éditeurs français du XIXe siècle*, projet d'ouverture et enrichissement d'une base de donnée des éditeurs en France au XIXe siècle, a d'abord été créé entre 2014 et 2015 au format MySQL en local sur les serveurs de l'Université de Versailles Saint-Quentin-en-Yvelines. Son modèle de données a néanmoins été transformé en 2016 lors de l'hébergement par la TGIR Huma-Num pour adopter les principes du web sémantique¹⁹, jugé « une voie d'avenir²⁰ ».

En tout état de cause, cet état de l'art de l'encodage des dictionnaires en XML/TEI nous permet de conclure que :

- La structuration sémantique classique des dictionnaires, thésaurus et documents structurés par des entrées se base sur une dichotomie entre la forme prise par le concept (le libellé de l'entrée) et son sens (bloc définitionnel). Ce phénomène est observable et facilement encodable dans notre projet.
- L'encodage en XML/TEI des dictionnaires se constitue principalement de nu-

18. Système de gestion de base de données fondé sur la langage XML.

19. Le web sémantique ou web de données est un modèle de donnée assurant leur échange et leur interopérabilité en se basant sur le *Resource Description Framework*.

20. Jean-Charles Geslot et Viera Rebolledo-Dhuin, « Du livre imprimé au Web sémantique : le projet du Dictionnaire des éditeurs français du xixe siècle », *Humanités numériques*-2 (1^{er} juin 2020), DOI : 10.4000/revuehn.426, p.12.

mérisation de dictionnaires déjà existants et rarement de documents nativement numériques comme le nôtre.

- Les domaines de la linguistique se sont très tôt emparés de la question de l'encodage en TEI, proposant dès la P3 un ensemble de balises traitant de l'étymologie et des aspects grammaticaux. Ces modèles étant peu utiles pour le projet FermeGé, il nous faut ouvrir notre champs de réflexion vers bases de données et vaste corpus de documents.
- Parmi ces différentes bases de données, le format XML/TEI est considéré comme un modèle de données sémantique, ainsi qu'un moyen de pérennisation des données en raison de la stabilité du schéma TEI (P5 sortie en 2007).
- La question de la structuration globale du dictionnaire se pose : faut-il proposer un encodage donnant lieu à un document, certes unique, mais massif ou passer par un encodage avec une balise `<teiCorpus>` ?

Chapitre 3

Une approche par la pratique. Premier encodage manuel prospectif et enjeux de modélisation

Afin de terminer cette analyse des enjeux et problématiques scientifiques ou techniques liés au *Dictionnaire numérique de la Ferme générale*, il nous semble essentiel de mettre en lumière les premières conclusions que nous avions pu tirer à l'issue d'un premier encodage manuel et prospectif de certaines notices. En effet, à partir de cet encodage, nous avons élaboré un modèle conceptuel de la structuration des données du Dictionnaire. Nous comprenons le terme de « modèle conceptuel » comme la représentation et schématisation des données sous la forme d'entités, caractérisées par des attributs et des logiques de relations. Dans notre cas, ce modèle conceptuel se traduit par une imbrication des blocs structurels et sémantiques. A partir de ce modèle conceptuel, nous avons pu mettre en oeuvre un « modèle logique », à savoir l'adaptation du modèle conceptuel dans un langage formel de structuration des données, en l'occurrence la TEI. Cette première approche de l'encodage nous permet de questionner et conceptualiser la structuration logique et sémantique du *Dictionnaire*. Quels sont les enjeux principaux de modélisation des données propres au Dictionnaire de la Ferme générale ? En quoi ce modèle conceptuel place le Dictionnaire à la confluence des bases de données textuelles et documents structurées par des entrées ?

Nous proposons d'analyser trois aspects essentiels du modèle d'encodage du *Dictionnaire* : le modèle conceptuel global, donnant lieu à une réflexion sur l'articulation des notices à l'échelle du dictionnaire, puis de l'encodage des entités nommées au sein des notices. Soulignons que ce présent chapitre se lit en miroir du suivant. Nous analysons ici les enjeux théoriques et conceptuels pour mieux mettre en perspective les choix techniques que nous développons dans le chapitre 4. Ce travail s'inscrit dans la seconde étape de notre *workflow* que nous schématisons ci-dessous :

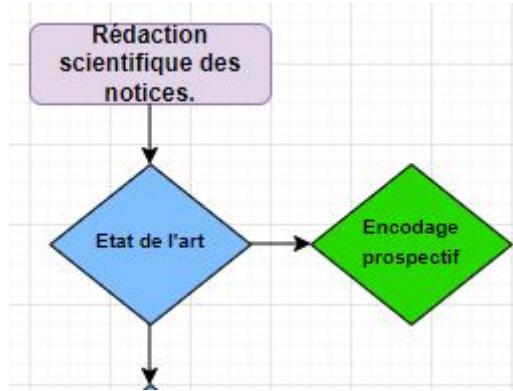


FIGURE 3.1 – *Workflow* : phase d’encodage manuel et prospectif

3.1 Rendre le sens explicite : vers une modélisation de la structure

3.1.1 La structuration conceptuelle des notices

Dans un premier temps, comme nous l’avons mentionné précédemment, le *Dictionnaire de la Ferme générale* s’inscrit non seulement dans la lignée des dictionnaires encyclopédiques visant à proposer une analyse d’histoire totale, mais aussi dans une perspective de « base de données » documentaire sur un modèle « clé-valeur ». C’est pourquoi nous proposons de mettre en place un encodage des notices reprenant le principe de la dualité sémantique entre la forme prise par le concept (le lemme) et son sens développé dans la définition scientifique. En somme, en nous basant sur une notice extraite du document de travail, la structure sémantique du Dictionnaire et de ses notices peut être schématisée et conceptualisée de la manière suivante :

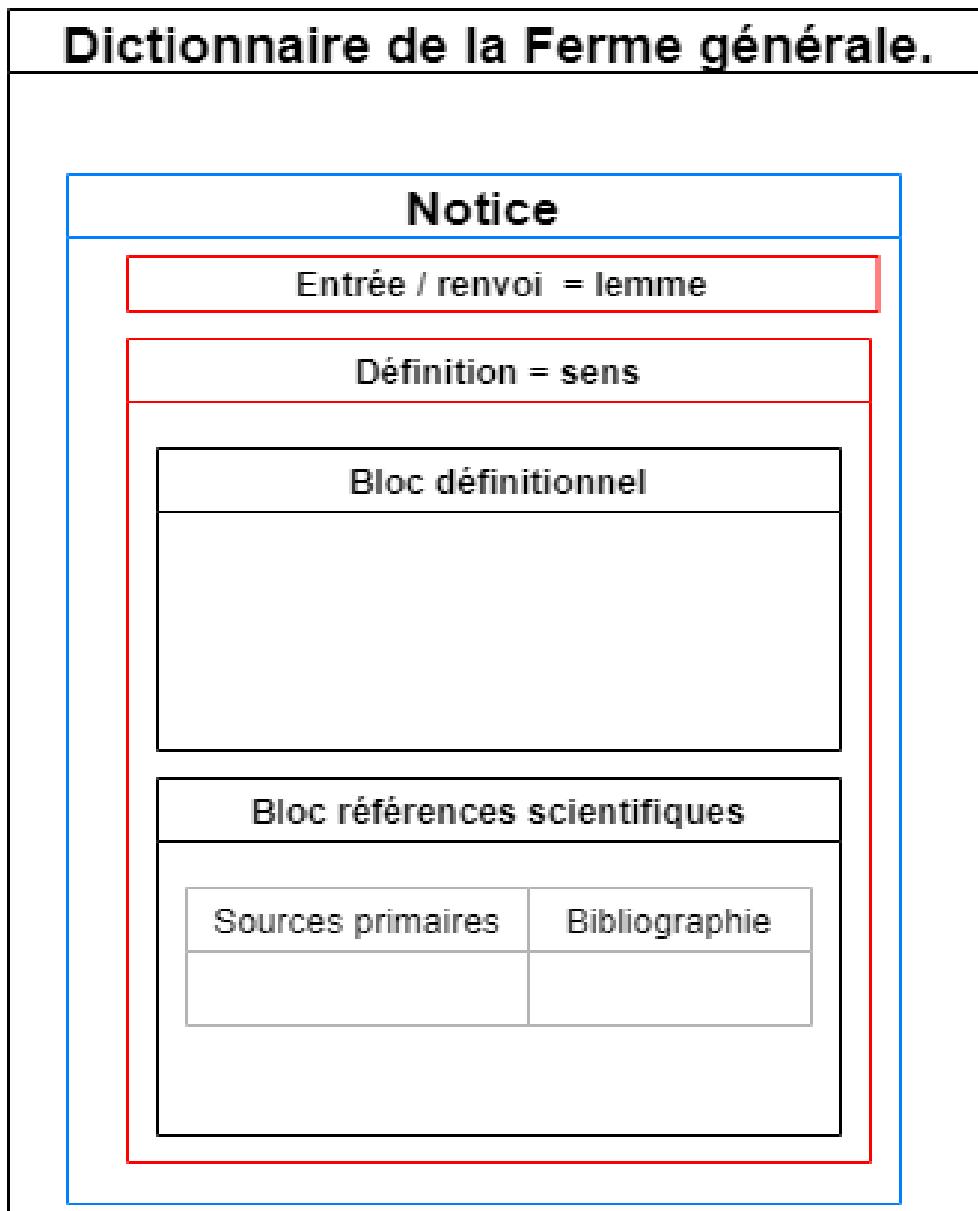


FIGURE 3.2 – Schéma du « modèle conceptuel » du Dictionnaire de la Ferme générale.

Le dictionnaire est l'entité englobante regroupant une collection de notices. À chacune de ses notices correspond donc une entrée englobante qui comporte d'une part le lemme du concept énoncé, et d'autre part la définition scientifique qui lui est associée. Cette définition se sépare en deux « blocs » sémantiques : le « bloc définitionnel » qui propose une définition et analyse scientifique du concept, et le « bloc des références scientifiques ». Dans la tradition des études historiques françaises, nous proposons de séparer conceptuellement ces données entre deux sous-ensembles : la liste des sources primaires mobilisées et la bibliographie scientifique. Soulignons donc que cette imbrication des éléments sémantiques nous conforte dans notre choix d'un encodage en XML/TEI.

3.1.2 Vers une esquisse d'un modèle logique

En conséquence, nous pouvons proposer une adaptation de ce modèle concept en un modèle logique impliquant un encodage en XML/TEI. Le schéma suivant indique les correspondances entre les différentes entités et les balises correspondantes :

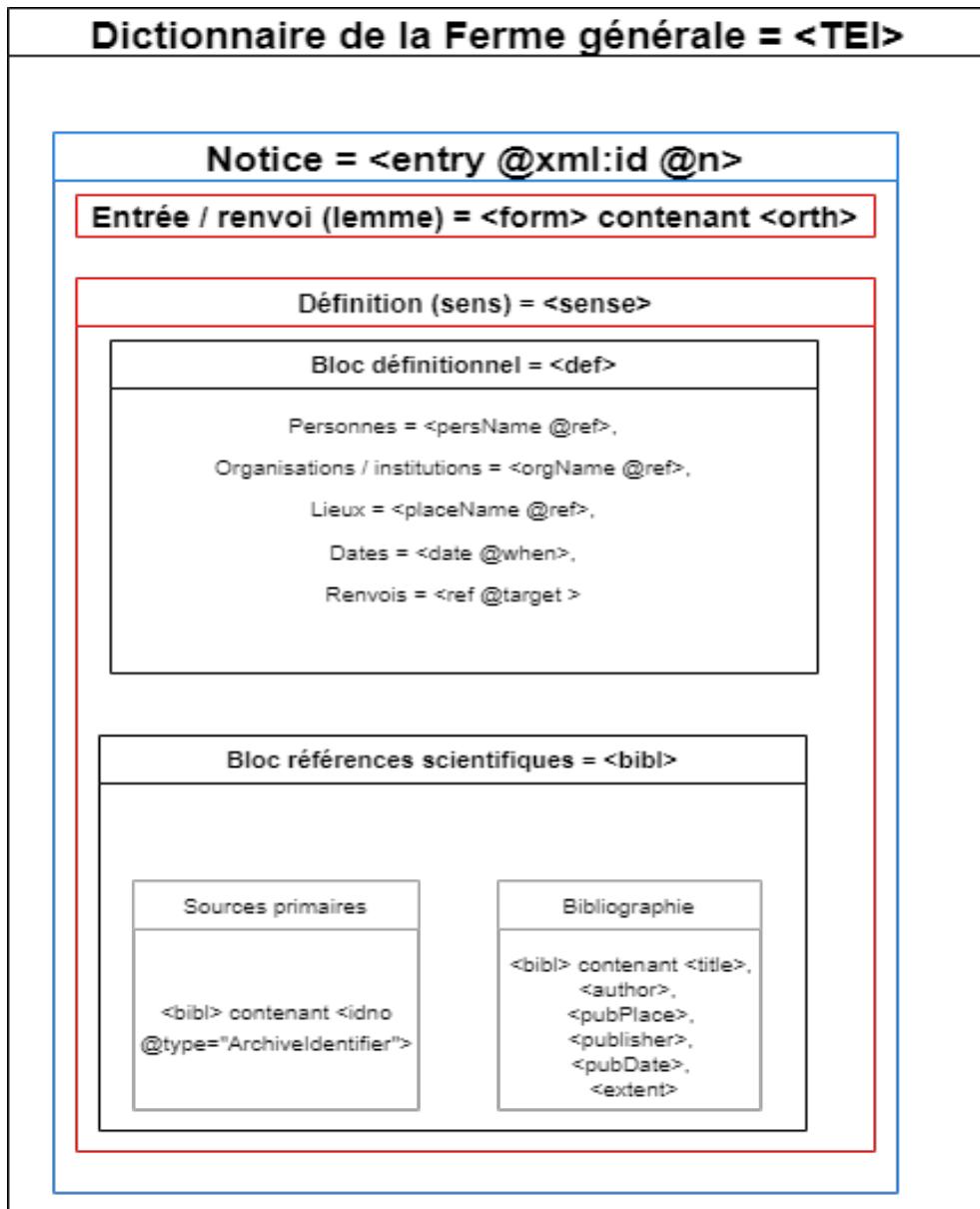


FIGURE 3.3 – Schéma du « modèle logique » du *Dictionnaire de la Ferme générale*.

Ce modèle logique se traduit par une implémentation physique avec l'extrait de l'encodage minimal d'une notice suivant, tiré du logiciel Oxygen XML Editor :

```

-----+
<entry xml:id="" p="">
  <form type="lemma">
    <orth>

      </orth>
    </form>
    <sense>
      <def>
        <bibl>
          <bibl><idno></idno></bibl>
          <bibl> <author></author><title></title><pubPlace></pubPlace>
                  <publisher></publisher><date></date><extent></extent>
            </bibl>
          </bibl>
        </def>
      </sense>
    </entry>

```

FIGURE 3.4 – Encodage XML/TEI minimal de la structure d'une notice.

3.1.3 Conceptualiser les liens entre les notices

Cet encodage structurel nous semble être un compromis intéressant entre la tradition de l'encodage des dictionnaires en XML/TEI développée par les linguistes et l'encodage des corpus de données historiques que nous avons précédemment cité. Dans la perspective de création d'index, d'intégration dans une application web et de mise en relation des notices par le biais des hyperliens, nous proposons d'ajouter à la balise englobante `<entry>` un attribut `@xml :id` permettant d'indexer et de renvoyer vers la notice en question. Un second attribut `@n` permet par la suite l'indexation des entités nommées. Si du point de vue de la structure des notices l'encodage semble relativement clair et facile à mettre en place, son articulation à l'échelle de l'ensemble du dictionnaire d'une part, et des entités ou « blocs » englobés d'autre part, est plus complexe et nous permet de soulever différentes hypothèses que nous devons à présent étudier.

3.2 L'articulation des notices à l'échelle du dictionnaire : analyse et évaluation des hypothèses

3.2.1 *Le Dictionnaire de la Ferme générale* : un corpus de notices scientifiques ?

Tout d'abord, à l'échelle du Dictionnaire dans sa globalité, deux hypothèses s'offrent à nous : soit encoder l'ensemble des notices dans un seul et même document XML/TEI

dans lequel à chaque notice correspond une balise <entry> sur le modèle précédemment exposé ; soit encoder chaque notice à part entière dans un document XML/TEI et les relier dans un document parent englobant par le biais d'une balise <teiCorpus>. Les modèles impliquent des choix scientifiques et conceptuels différents. L'hypothèse de l'encodage « en <teiCorpus> » est intéressante car elle permet une segmentation du corpus des notices renforçant leur unité intellectuelle propre. De plus, d'un point de vue technique, cela autorise l'encodage d'un <teiHeader> accueillant les métadonnées et index par notice. Ce modèle d'encodage a notamment été employé pour le projet ThEMA, utilisant la déclaration d'un espace de nom « xmlns : xi » et des éléments <xi :include>. L'extrait suivant du « fichier maître » du projet ThEMA¹ illustre le processus d'encodage adopté :

```
<teiCorpus xmlns="http://www.tei-c.org/ns/1.0" xmlns:xi="http://www.w3.org/2001/XInclude"
           xmlns:dc="http://dublincore.org/documents/dcmi-namespace/" type="collection" xml:id="TC0001">
  <teiHeader>
    <fileDesc>
      <titleStmt>
```

FIGURE 3.5 – Extrait de l'encodage du fichier teiCorpus du projet ThEMA.

Ce choix effectué au sein du projet ThEMA se justifie par le fait qu'une grande partie du contenu de cette base de données est constituée des métadonnées des différents *exempla*, se trouvant donc encodés dans les <teiHeader²> correspondant.

3.2.2 Le *Dictionnaire de la Ferme générale* : une œuvre organique ?

Si cette première hypothèse est intéressante pour le Dictionnaire, elle présente néanmoins certaines limites conceptuelles et techniques. Effectivement, elle implique le fait que le Dictionnaire soit considéré comme une collection ou un corpus de notices indépendantes. Or, le *Dictionnaire de la Ferme générale* n'est-il pas une entité à part entière et ne présente-t-il pas une unité conceptuelle et intellectuelle ? Dès lors, l'hypothèse d'un encodage par le biais d'un seul document englobant, où chaque notice correspond à une balise <entry> doit être évaluée. Cette solution nous semble jusqu'à présent la plus viable d'un point de vue structurel. Effectivement, cela nous permet de fournir un <teiHeader> unique présentant les métadonnées pour l'ensemble du Dictionnaire, considéré comme une œuvre scientifique à part entière. De même, cette solution autorise la création des divers index des noms de personnes, de lieux, d'organisations et institutions concernant l'ensemble des notices. La problématique sous-jacente relève donc de l'aspect technique

1. J.P. Rehr et M.A. P. d. Beaulieu, « Thesaurus Exemplorum Medii Aevi... ».

2. Balise TEI renseignant les métadonnées descriptions et d'indexation du document. Voir <https://tei-c.org/release/doc/tei-p5-doc/fr/html/ref-teiHeader.html>

de la manipulation, transformation et extraction des entités au sein d'un fichier unique et relativement volumineux. Finalement, au-delà de ces enjeux structurels, se pose la question de la place des entités nommées qui est centrale dans le cadre de l'élaboration du Dictionnaire.

3.3 Les entités nommées : encodage, extraction et mise à disposition

3.3.1 Les entités nommées : l'apanage de la linguistique ?

Le *Dictionnaire de la Ferme générale* entend répondre à un critère fondamental d'approche complète et non parcellaire, tant thématique qu'analytique. C'est pourquoi il nous semble pertinent d'en proposer un enrichissement par le biais de différents index recensant les entités nommées. Ces index concernent les noms de lieux et d'organisation. Il nous est de même possible de fournir une chronologie pour appuyer le travail de l'équipe scientifique. L'objectif de cette indexation est de pouvoir faciliter la navigation au sein du corpus des notices pour la communauté scientifique, et ainsi que de créer des liens entre l'axe 1 du projet et son axe 4 (« ancrage dans les principes de la science ouverte »).

Afin de saisir les enjeux scientifiques et techniques de cette indexation, il nous faut d'abord rappeler ce que sont les entités nommées et en quoi leur traitement est un enjeu à part entière. Il s'agit de l'ensemble des expressions linguistiques référentielles, comprenant les anthroponymes (entités humaines historiques), toponymes (localisable géographiquement) et les ergonymes (organisations, institutions, gouvernements et sociétés³). Diverses réflexions et débats, notamment au sein du Consortium de la TEI ont permis d'inclure dans ce concept les entités dites « abstraites », faisant références à des personnalités morales, des personnages pseudo-historiques ou des allégories et divinités clairement nommées⁴. Le traitement et l'analyse des entités s'inscrit dans un champ de recherche vaste relevant du traitement automatique des langues (« Natural Language Processing »), à la frontière de la linguistique, de l'intelligence artificielle et de l'étude historique du langage. Ce domaine interdisciplinaire entend analyser le langage naturel (en opposition au langage informatique) par le biais des outils numériques de traitement automatique. Ainsi, il recouvre, en amont, la lemmatisation, la morphologie, l'étiquetage morpho-syntaxique, l'analyse syntaxique et sémantique (« Parts of Speech »), la « racinisation » ou « désuffixation »

3. Damien Nouvel, *Reconnaissance des entités nommées par exploration de règles d'annotation : interpréter les marqueurs d'annotation comme instructions de structuration locale*. These de doctorat, Tours, 2012, URL : <http://www.theses.fr/2012TOUR4011> (visité le 14/04/2022).

4. Daniel Jurafsky et James H. Martin, *Speech and Language Processing. An introduction to Natural Language Computational Linguistics, and Speech Recognition*, Third Edition draft, 2022, URL : https://web.stanford.edu/~jurafsky/slp3/ed3book_jan122022.pdf, p.25.

(« tokenization ») ainsi que la séparation automatique des mots (« parsing »). En aval se trouvent la fouille de texte, classification des documents, reconnaissance optique de caractère (« Ocular Character Recognition »), d'écriture manuscrite (« Handwritting Text Recognition ») et la reconnaissance d'entités nommées. Dans ce réseau dense et interdisciplinaire, le traitement et l'extraction des entités nommées présente des difficultés propres auxquels nous sommes directement confrontés. La première est l'ambiguïté de la sémantique d'une entité nommée⁵. Effectivement, soulignons tout d'abord qu'il apparaît particulièrement complexe pour un système informatique de reconnaître un nom de lieu ou de personne et tout particulièrement de les différencier et classer dans leurs catégories respectives d'anthroponyme ou de toponyme. La seconde est l'ambiguïté de segmentation liée aux noms, tout particulière lorsqu'il s'agit de noms composés ou références à une entité spécifique.

3.3.2 Etat de l'art du traitement des entités nommées dans les sciences historiques

L'étude, la reconnaissance et l'extraction des entités nommées dans le domaine des humanités numériques appliquées à l'Histoire se sont fortement développées depuis le milieu des années 2000 en lien avec un processus de numérisation massive des documents. Les projets d'encodage ont permis d'établir progressivement des processus de traitement de la donnée sur des documents divers tant dans leur sujet que dans leur support. Dès 2006, le travail de reconnaissance des entités nommées appliqué aux journaux américains, mené par les chercheurs G. Crane et A. Jones, met en exergue la possibilité de constituer des corpus massifs d'entités nommées⁶. Ce projet ouvrit donc la voie à de nombreux champs d'application tels que, les documents d'archives⁷, des documents numérisés⁸ ou encore des documents traitant de catégories sémantiques allégoriques (par exemple, l'animisme⁹ au sens de l'*animacy* anglo-saxonne). Une phase de ce développement de la reconnaissance des entités nommées est indéniablement l'ouverture des corpus et méthodes de traitements aux documents et langues extra-européennes. A cet égard citons les travaux de Noriyoshi

5. *Ibid.*, p.79.

6. Gregory Crane et Alison Jones, « The challenge of virginia banks : an evaluation of named entity analysis in a 19th-century newspaper collection », dans *Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries*, New York, NY, USA, 2006 (JCDL '06), p. 31-40, DOI : 10.1145/1141753.1141759.

7. Kate Byrne, « Nested Named Entity Recognition in Historical Archive Text », dans *School of Informatics*, University of Edinburgh, 2007, p. 589-596, DOI : 10.1109/ICSC.2007.107.

8. M. Khemakhem, Luca Foppiano et L. Romary, « Automatic Extraction of TEI Structures in Digitized Lexical Resources using Conditional Random Fields », dans *electronic lexicography, eLex 2017*, Leiden, Netherlands, 2017, URL : <https://hal.archives-ouvertes.fr/hal-01508868> (visité le 12/04/2022).

9. L. Borin, D. Kokkinakis et Leif-Jöran Olsson, « Naming the Past : Named Entity and Animacy Recognition in 19th Century Swedish Literature », dans *LaTeCH@ACL*, 2007.

Nagi, Fuminori Kimuro, Akira Maeda et Ryo Akama concernant les entités nommées dans les documents historiques japonais¹⁰.

En outre, les progrès en matière de reconnaissance des entités nommées ces dernières années permettent une ouverture des données et un croisement avec d'autres types et structuration de l'information, notamment des données géographiques. Comme le mentionne les chercheurs Miguel Won, Patricia Murrieta-Flores et Bruno Martins¹¹, ce croisement des informations géographiques avec les entités nommées dans le cadre des toponymes et documents historiques pose encore divers problèmes. Tout d'abord, l'enjeu de désambiguïsation est d'autant plus prégnant dans le cadre des toponymes historiques que la graphie peut varier d'une époque à l'autre, et recouvrir des réalités sensiblement distinctes des dénominations contemporaines. En outre, une institution historique peut se confondre avec le territoire qu'elle contrôle, administre ou protège. Dès lors, l'encodage doit prendre en compte cette proximité et proposer un modèle incluant non seulement l'organisation ou le toponyme mentionné. En guise de synthèse, le schéma suivant entend synthétiser les domaines disciplinaires auxquels nous devons nous référer pour le traitement des entités nommées :

10. Noriyoshi Nagai, Fuminori Kimura, Akira Maeda et Ryo Akama, « Personal Name Extraction from Japanese Historical Documents Using Machine Learning », dans *International Conference on Culture and Computing*, 2015, p. 207-208, DOI : 10.1109/Culture.and.Computing.2015.46.

11. Miguel Won, Patricia Murrieta-Flores et Bruno Martins, « Ensemble Named Entity Recognition (NER) : Evaluating NER Tools in the Identification of Place Names in Historical Corpora », *Frontiers in Digital Humanities*, vol. 5 (2018). URL : <https://www.frontiersin.org/articles/10.3389/fdigh.2018.00002>. Consulté le 16 août 2022.

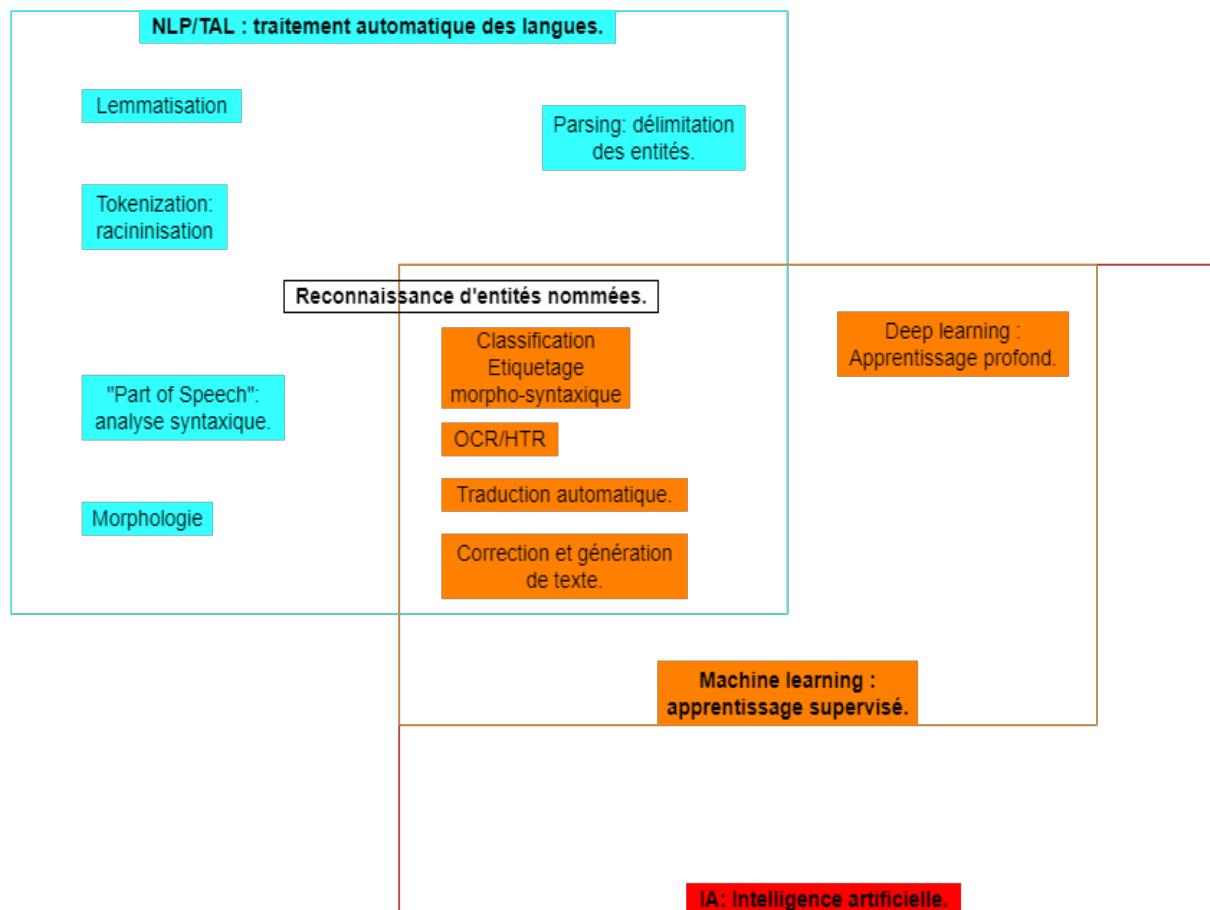


FIGURE 3.6 – Diagramme la reconnaissance d’entités nommées dans les différentes approches des humanités numériques (TAL/NLP)

3.3.3 Les entités nommées et le projet FermeGé : quelles perspectives ?

Dans le cadre du projet, l’encodage des entités nommées de personnes peut présenter une différence de granularité parfois significative. Effectivement, nous sommes confrontés à la possibilité d’encoder les noms de personnes comme des entités unitaires sur le modèle suivant :

```
<persName ref="#Jacques_Necker">Jacques Necker</persName>
```

FIGURE 3.7 – Capture d’écran : proposition d’encodage des anthroponyme sous forme d’entités unitaires.

Nous pouvons cependant gagner en finesse dans la granularité en précisant qu’une entité nommée contient un nom de famille et prénom :

```
<persName ref="#Jacques_Necker"><forename>Jacques</forename> <surname>Necker</surname></persName>
```

FIGURE 3.8 – Capture d’écran : proposition d’encodage des anthroponymes avec une granularité fine.

Cette possibilité d’encodage pose néanmoins la question de la régularité et l’uniformité des entités nommées de personnes puisque certaines sont des références à des personnes morales sans être explicitement nommées comme dans le cas suivant :

```
<persName ref="#Duc_de_Bretagne">duc de Bretagne</persName>
```

FIGURE 3.9 – Capture d’écran : proposition d’encodage des entités morales.

Afin d’assurer une pérennisation, ouverture et réutilisation des données, il nous faut déterminer une ontologie des références des entités nommées de personnes. Il nous semble envisageable dans la temporalité plus longue du projet de fonder cette ontologie sur les notices d’autorité du *Virtual International Authority File* (VIAF) qui, grâce aux liens pérennes et aux identifiants proposés, permettent d’avoir accès à un référentiel partagé, ouvert et s’inscrivant dans les bonnes pratiques de l’interopérabilité des données. Nous pouvons donc adapter notre modèle d’encodage à ce choix comme le montre l’exemple suivant¹² :

```
<persName ref="http://viaf.org/viaf/68949881">Jacques Necker</persName>
```

FIGURE 3.10 – Capture d’écran : proposition d’encodage de ”persName” avec insertion du référentiel VIAF.

Une réflexion identique doit s’appliquer aux entités nommées d’organisations et d’institutions. Nous proposons dans ce cas d’encoder dans une balise *<orgName>* l’ensemble des institutions politiques, financières, judiciaires et militaires mentionnées dans les notices afin de pouvoir en tirer un index. Cet encodage soulève néanmoins des questions d’ordre scientifique lorsqu’il s’agit d’indexer des institutions ayant une forte emprise géographique. Faut-il encoder par exemple la « généralité d’Amiens » comme une entité nommée de lieu ou comme une institution ? Il semble que dans les cas des états provinciaux et des généralités, cette « organisation » soit avant tout une réalité politique qu’il conviendrait d’encoder avec des balises *<orgName>*.

```
<orgName ref="#Généralité_Amiens">généralité d'<placeName ref="#Amiens">Amiens</placeName></orgName>
```

FIGURE 3.11 – Capture d’écran : proposition d’encodage de ”orgName”

12. Notons que l’automatisation de la récupération de ces notices d’autorité peut être mise en place par le requêtage de l’API de la BnF disponible à cette adresse : <https://api.bnf.fr/fr/notices-dautorite-personnes-collectivites-oeuvres-lieux-noms-communs-de-bnf-catalogue-general>

De même, la plupart de ces institutions et organisations sont associées à un territoire plus ou moins finement localisé. C'est pourquoi la question de la granularité de l'indexation se pose. En effet, devons-nous encoder uniquement « généralité d'Amiens » comme une institution à part entière ou préciser l'encodage, en ajoutant l'indexation de l'entité nommée de lieu ? Il nous semble pertinent de proposer cet encodage imbriqué dans la mesure où il permet l'ajout d'une couche sémantique supplémentaire et d'une double indexation. Il nous faut cependant croiser ce modèle avec la procédure de cartographie inhérente à la réalisation de l'Atlas (axe 2 du projet FermeGé) et au choix de l'échelle de localisation des différentes entités spatiales. La même problématique de l'ontologie se pose et du possible ajout de coordonnées géographiques au sein des balises ou dans l'index. Notons de même le cas de l'encodage des renvois vers d'autres notices marquées typographiquement dans le document de travail par des astérisques accolées aux mots (ex : « aides* »). Ce procédé a permis dans un premier temps aux rédacteurs des notices de baliser des termes faisant ou devant faire l'objet d'une notice à part entière. Dès lors, nous proposons d'encoder en TEI ces renvois par la balise <ref> complétée d'un élément @target. A terme, ce balisage permettra d'inclure un lien renvoyant vers la notice mentionnée.

*

Au cours de cette première partie contextualisante et prospective nous avons pu identifier les enjeux et caractéristiques majeures associées au *Dictionnaire de la Ferme générale* à savoir :

- Le caractère nativement numérique des sources que nous avons à notre disposition, à savoir un ensemble de notices extraites du carnet Hypothèses.org du projet et rassemblées dans un document texte. Les enjeux sont donc double puisqu'il nous faut proposer une chaîne de traitement de la données afin de la rendre interopérable et pérennisable.
- Une structuration conceptuelle qui invite à penser le dictionnaire à la fois comme une collection de notices individuelles et comme une oeuvre cohérente. Il est donc nécessaire de proposer un schéma d'encodage faisant un choix entre ces deux alternatives en cohérence avec les enjeux scientifiques du projet.
- La nécessité de proposer une mise à disposition des notices et des données de la recherche dans le cadre notamment d'une application web permettant notamment la visualisation des notices et la navigation dans un index des entités nommées.

*

Deuxième partie

**Du schéma à l'automatisation de
l'encodage : mise en oeuvre d'une
chaîne de traitement de la donnée**

Chapitre 4

Structurer, schématiser et modéliser le *Dictionnaire de la Ferme générale* : l’O.D.D et son application

A la suite de l’encodage manuel et prospectif des « notices tests », nous avons donc pu déterminer un modèle conceptuel et logique d’encodage des données du Dictionnaire. Ce chapitre se veut donc le versant technique du précédent puisque nous détaillons ici les choix et solutions mises en place. Afin de pouvoir d’une part formaliser un ensemble de règles propres au projet du *Dictionnaire numérique de la Ferme générale*, et d’autre part d’assurer la mise en place d’une chaîne de pose automatique de balises TEI, il nous faut rédiger un O.D.D. Il s’agit d’un document biparti contenant dans une première section la documentation « en prose » du schéma d’encodage et de validation, puis dans une seconde le schéma en lui-même, rédigé dans le langage même de la TEI. L’objectif principal d’un O.D.D (*One Document Does it all*) est donc de restreindre l’utilisation de la TEI dans le cadre du projet à un certain nombre de balises et de situations. Ce choix de restreindre la TEI à un schéma spécifique ne doit nullement surprendre le lecteur. En effet, comme le rappel Lou Burnhard, co-fondateur du *TEI Consortium* dans son ouvrage *Qu’est-ce que la TEI?*, les objectifs principaux de ce métalangage sont l’échange et le partage des documents encodés¹. Dès lors, afin de rendre les documents TEI réutilisables par la communauté des chercheurs, et par extension pérennisables, un schéma d’encodage clair et documenté est nécessaire. Indiquons le fait que plusieurs langages de schémas existent pour structurer et restreindre l’utilisation des langages fondés sur l’*eXtensible Language Markup* (XML) :

- la D.T.D ou *Document Type Definition*².

1. L. Burnard, *Qu’est-ce que la Text Encoding Initiative ?...*, Voir Chapitre 7 « Personnaliser la TEI ».

2. w3school, DTD Tutorial, url : https://www.w3schools.com/xml/xml_dtd_intro.asp

- les schémas XML basés sur les recommandations du W3C : *W3C Schema*³.
- le RelaxNG (RRegular LAnguage for XML Next Generation)⁴.
- le Schematron⁵.

Notre attention s'est portée sur la rédaction d'un schéma dans la « grammaire » O.D.D, produisant un document au format XML et ensuite convertit au format RNG (RelaxNG) grâce aux moteurs et feuilles de transformation internes au logiciel Oxygen. En somme, le fichier ODD contient une liste des éléments autorisés, leur contexte d'utilisation, leur attributs et valeurs requises ou permises, ainsi qu'un ensemble d'exemples (personnalisés ou générés à partir des *Guidelines*). De plus, le recours à la grammaire de l'O.D.D permet d'insérer des règles contextuelles plus précises en langage *Schematron* après avoir déclaré cet espace de noms dans le préambule du document. Dans notre cas, l'O.D.D doit donc répondre à la nécessité des enjeux scientifiques du projet tout en prenant en compte l'ensemble de la chaîne technique de traitement et de balisage automatique. Il nous faut donc nous questionner sur la mise en adéquation des besoins scientifiques du modèle conceptuel précédemment énoncé avec les enjeux et contraintes techniques de l'O.D.D. Dans quelle mesure la mise en oeuvre du schéma d'encodage nourrit réciproquement le modèle conceptuel et théorique ?

Pour ce faire, nous proposons d'appuyer notre démonstration sur des exemples révélateurs des enjeux et tensions liés aux choix que nous avons opérés, plutôt que de proposer une gloser exhaustive de notre O.D.D⁶. Cette analyse s'inscrit dans la seconde phase majeure de notre *workflow* que nous complétons ci-dessous :

3. <https://www.w3.org/standards/xml/schema.html>

4. RelaxNG, *Home*, relaxng.org, <https://relaxng.org/>

5. Schematron, *Home*, schematron.com, url : <http://schematron.com/>

6. La documentation exhaustive et le code commenté de l'O.D.D sont à cet égard disponibles dans les fichiers joints à ce présent mémoire (voir annexes B.2) ou à l'adresse suivante : https://gitlab.huma-num.fr/vdecreaene/DicoNumFermeGe/-/tree/main/ODD_schemas

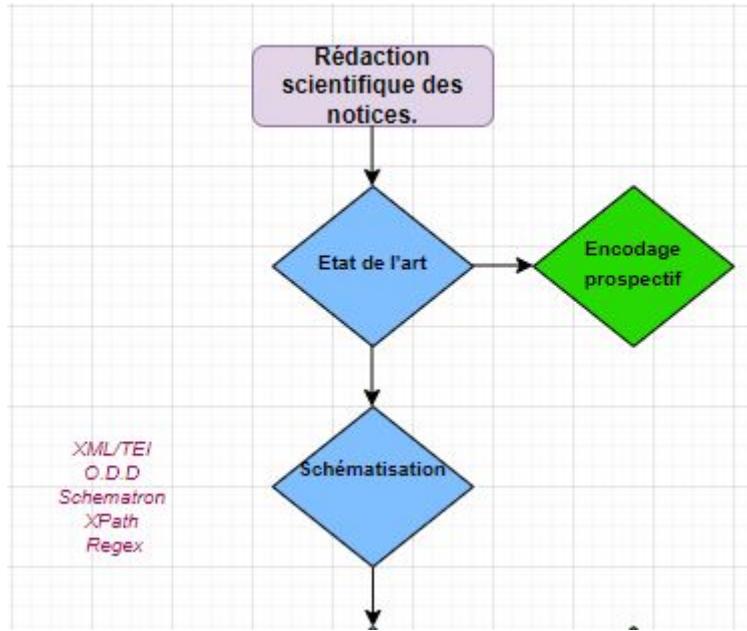


FIGURE 4.1 – *Workflow* : Schématisation et conceptualisation de l’O.D.D

4.1 L’O.D.D des notices individuelles : le Dictionnaire comme une collection d’entités distinctes

Dans un premier temps, nous avons rédigé et formalisé un O.D.D en considérant le dictionnaire comme une collection de notices individuelles bien que liées entre elles par de nombreux renvois et des thématiques communes. Ce choix de schématisation est basé sur l’établissement du modèle conceptuel que nous avions pu établir lors de l’encodage prospectif dont nous proposons ici pour rappel une visualisation par arborescence, afin d’établir un lien direct avec le modèle logique de données par arbres XML :

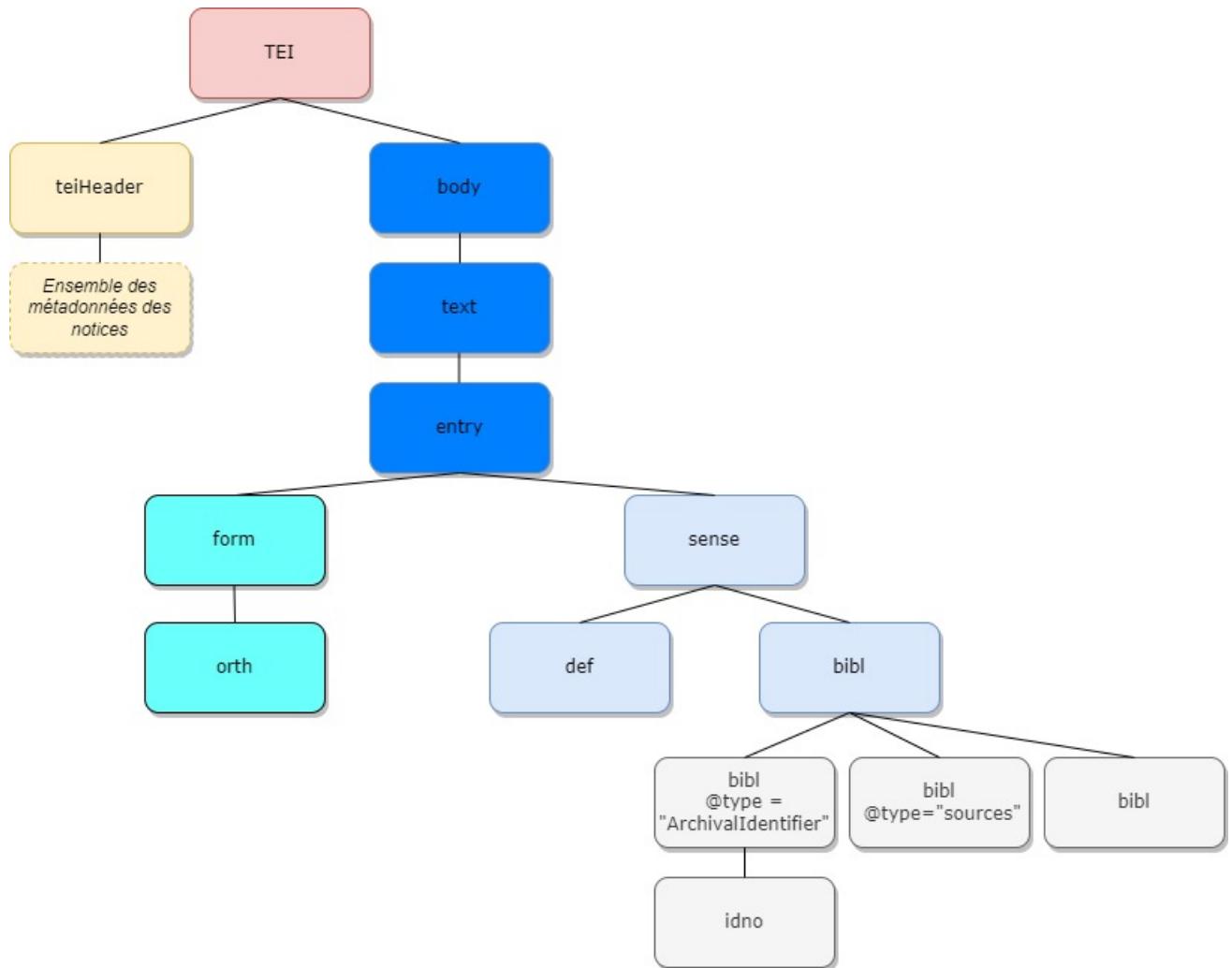


FIGURE 4.2 – Schéma : modèle logique en arborescence XML/TEI des notices.

4.1.1 La schématisation de la structure des notices : le module *dictionnaires* et ses limites

D'un point de vue structurel, nous avons donc eu recours au module *dictionnaires* de la TEI. Rappelons que la TEI doit son efficacité et sa flexibilité à son caractère modulable. Quatre modules sont obligatoires : *tei* (définissant les classes, macros et types de données), *core* (contenant les éléments disponibles dans tous les documents), *header* (permettant de renseigner les métadonnées dans l'en-tête) et *textstructure* (incluant les éléments de structure des documents textuels). Le module *dictionnaires* est quant à lui un module additionnel que nous avons donc dû déclarer dans le schéma⁷ comme le montre l'extrait ci-dessous. Nous ne commentons dans ce chapitre que les exemples les plus révélateurs des choix effectués. L'ensemble de l'O.D.D commenté est trouvable en annexe A7.3.3

7. Afin de gagner en clarté et de s'assurer la conformité la plus stricte du schéma aux besoins du projet, nous avons décidé de rédiger l'O.D.D manuellement. Les modules sont déclarés dans une partie dédiée du schéma, les uns à la suite des autres. Indiquons qu'il est possible de générer automatiquement des O.D.D grâce à des feuilles de transformation interne au logiciel Oxygen tel qu'*ODD by example*

```
<moduleRef key="namesdates" include="persName orgName placeName org"/>
<moduleRef key="dictionaries" include="entry form orth sense def"/>
<moduleRef key="figures" include="figure graphic"/>
```

FIGURE 4.3 – Capture d’écran : déclaration des modules additionnels « namesdates », « dictionaries » et « figures ».

Soulignons d’ores et déjà que nous avons utilisé une quantité restreinte de balises disponibles dans ce module, à savoir les éléments `<entry>`, `<form>`, `<orth>`, `<sense>` et `<def>`, afin d’encoder la structure logique. Effectivement, l’ élément englobant `<entry>` comprend l’ensemble d’une notice de dictionnaire, séparée en deux sous-ensembles majeurs : le lemme contenu dans l’élément `<form>` et le sens dans l’élément `<sense>`. Au sein de ce dernier, nous avons décidé de schématiser une séparation entre la définition (`<def>`) et la bibliographie (`<bibl>`, que nous analysons plus loin). Le cas des règles concrètes de schématisation de la balise `<entry>` nous semble particulièrement intéressant dans la mesure où il combine la grammaire O.D.D et le langage *Schematron*⁸ pour préciser le contexte d’utilisation de certains attributs :

```
<elementSpec ident="entry" mode="change">
  <gloss>Entrée structurée de dictionnaire.</gloss>
  <desc>contient le corps du texte de la notice encodée (lemme, définition,
    références).</desc>
  <content>
    <sequence preserveOrder="true">
      <elementRef key="form"/>
      <elementRef key="sense"/>
    </sequence>
  </content>
  <constraintSpec scheme="schematron" ident="entry">
    <constraint>
      <s:rule context="//tei:entry">
        <s:assert test="@xml:id"> L'élément entry doit contenir
          un attribut @xml:id </s:assert>
      </s:rule>
      <s:rule context="//tei:entry[@xml:id]">
        <s:assert test="matches(@xml:id, '^\\w+$')"> La valeur de
          l'attribut @xml:id doit contenir une suite de caractères ou
          de chiffres sans espaces </s:assert>
      </s:rule>
    </constraint>
  </constraintSpec>
</elementSpec>
```

FIGURE 4.4 – Capture d’écran : schéma O.D.D et *Schematron*, extrait du schéma appliqué à la balise `<entry>`

Effectivement, nous avons d’abord recours à la séquence d’éléments contenus dans la

8. Langage permettant de valider la structure d’un document XML en insérant des règles dites complexes touchant le contexte d’un élément ou son utilisation. Voir <https://schematron.com/>

balise <content⁹> pour déclarer les éléments autorisés au sein de la balise <entry¹⁰>, en l'occurrence <form¹¹> et <sense¹²>. Ensuite, nous utilisons le *Schematron* pour préciser les attributs autorisés et le type des valeurs de ces derniers grâce à une expression régulière et au langage Xpath¹³. La première règle *Schematron*, à savoir <s :rule context='//tei :entry[@xml :id]'> est un « test d'existence » et peut se traduire de la manière suivante : « afin d'être validé par le schéma, l'élément <entry> (tei :entry), descendant de l'élément racine (//) doit obligatoirement contenir un attribut xml :id ([@xml :id]) ».

La seconde règle concerne la valeur de cet attribut @xml :id qui doit correspondre à l'expression régulière indiquée, soit contenir une suite de caractères ou de chiffres sans espaces. Afin de fixer et restreindre la structure des notices, des règles comparables ont été formalisées pour les éléments <form>, <sense> et <def>. Nous avons utilisé la possibilité de restreindre l'ordre d'utilisation des balises afin d'assurer une validation uniforme de la structure, comme dans le cas de la balise <sense> :

```

<elementSpec ident="sense" mode="change">
    <gloss>Informations relatives au sens d'une entrée de dictionnaire</gloss>
    <desc>contient la définition de l'entrée de Dictionnaire et ses références
        scientifiques.</desc>
    <content>
        <sequence preserveOrder="true">
            <elementRef key="def"/>
            <elementRef key="bibl"/>
        </sequence>
    </content>
</elementSpec>
```

FIGURE 4.5 – Capture d'écran : règles de l'O.D.D portant sur l'ordre des balises au sein de la balise <sense>.

4.1.2 Les métadonnées et le teiHeader : de l'indexation à la pérennisation du projet

L'un des enjeux principaux du schéma a été de proposer une structuration des métadonnées des notices qui puisse à la fois répondre aux impératifs scientifiques, administratifs et techniques du projet. Dans cette optique, nous avons déclaré et utilisé les éléments suivants issus du module « header » :

9. Balise contenant la déclaration d'un modèle de contenu autorisé dans le cadre d'un schéma. Voir <https://tei-c.org/release/doc/tei-p5-doc/fr/html/ref-content.html>

10. Balise contenant l'entrée structurée d'un dictionnaire.

11. Balise encodant les informations relatives à la morphologie d'un mot. Voir <https://tei-c.org/release/doc/tei-p5-doc/fr/html/ref-form.html>

12. Balise regroupant les informations sur un terme, à savoir la définition et ses exemples. Voir <https://www.tei-c.org/Vault/P5/3.4.0/doc/tei-p5-doc/en/html/ref-sense.html>

13. Il s'agit d'un langage utilisé pour naviguer dans l'arborescence XML, voir https://www.w3schools.com/xml/xpath_intro.asp

```
<moduleRef key="header"
    include=" sourceDesc edition editionStmt keywords textClass language langUsage creation abstract
    extent sponsor funder fileDesc principal
    titleStmt publicationStmt profileDesc idno extent encodingDesc projectDesc revisionDesc
    change"/>
    ...
    "
```

FIGURE 4.6 – Capture d’écran : déclaration des éléments autorisés du module « header »

En premier lieu, les métadonnées descriptives concernant le projet et les informations administratives sont contenues dans la balise `<titleStmt>`, elle-même insérée dans la balise `<fileDesc>`. A cet égard, nous avons donc renseigné systématiquement les institutions en charge ou associées au projet. Nous proposons de séparer d’une part l’institution finançant le projet, identifiée par la balise `<funder>` (l’Agence Nationale de la Recherche) et d’autres part les institutions « partenaires », encodées avec les balises `<sponsor>` (le GIP Mission de recherche Droit et Justice, la MESHS, l’IRHiS, ect.). La personne ayant la responsabilité intellectuelle du projet et de sa coordination, madame Marie-Laure Legay est indiquée dans une balise spécifique (`<principal>`).

La plupart des métadonnées relevant de la description scientifique du projet et de ses enjeux historiographiques sont renseignées dans une balise `<projectDesc>`, contenue dans la balise `<encodingDesc>`. Nous avons ici repris la description scientifique proposée sur lors de la soumission du projet à l’ANR¹⁴.

Les métadonnées d’ordre technique sont indiquées dans diverses balises relevant à la fois de la conduite du projet, mais aussi de sa pérennisation sur le long terme. Effectivement, dès la schématisation et structuration des métadonnées, un impératif d’identification des données produites et de leur insertion dans une perspective de science ouverte s’est imposé à nous¹⁵. Cette mise en pratique de la science ouverte se traduit dans la gestion des métadonnées décrivant le processus et le contexte d’encodage. Il nous a semblé pertinent d’encoder dans une balise `<appInfo>` l’ensemble des informations relatives au logiciel d’encodage utilisé, en l’occurrence *OxygenXMLEditor* dans sa version 24.1. Cette décision s’inscrit dans la volonté de garder un historique du processus d’encodage du projet. Dans l’optique de la mise à disposition sur le long terme du Dictionnaire à la communauté des chercheurs, nous avons inclus dans notre schéma la possibilité de renseigner la licence ouverte qui pourra être appliquée. Par exemple, l’usage d’une licence Creative Commons (CC BY) est envisageable à terme dans le cadre de la diffusion des notices dans l’application web du *Dictionnaire Numérique de la Ferme générale*. En outre, nous avons proposé une « indexation matière » dans ce même `<teiHeader>`, au sein de la

14. https://anr.fr/fr/projets-finances-et-impact/projets-finances/projet/funded/project/anr-21-ce41-0019/?tx_anrprojects_funded%5Bcontroller%5D=Funded&cHash=5c64d94ca826534590e484c6f39658b300

15. Rappelons que l’un des enjeux de ce présent travail était de proposer des liens concrets et durables avec le 4e axe transversal du projet, à savoir la « pratique de la science ouverte »

balise <textClass>, proposant une liste de mots clés (balise <keywords>), à l'instar de l'extrait ci-dessous :

```

<textClass>
  <keywords scheme="#fr_RAMEAU">
    <list>
      <item> Histoire -- 17ème siècle -- Histoire des institutions </item>
      <item> Histoire -- 18ème siècle -- Histoire des institutions </item>
      <item> Histoire -- Dictionnaires </item>
      <item> Histoire -- Sources </item>
      <item> Histoire -- 17e siècle -- Histoire des doctrines </item>
      <item> Histoire -- 18e siècle -- Histoire des doctrines </item>
      <item> Histoire -- Historiographie </item>
      <item> Histoire -- Histoire religieuse </item>
      <item> Histoire -- 17ème siècle -- Histoire économique </item>
      <item> Histoire -- 18ème siècle -- Histoire économique </item>
      <item> Histoire -- Histoire sociale </item>
      <item> Histoire -- Histoire du droit </item>
      <item> Histoire -- Histoire judiciaire </item>
      <item> Histoire -- Histoire financière </item>
      <item> Histoire -- Impôt -- Administration et procédure </item>
      <item> Histoire -- Contestations </item>
      <item> Histoire -- Histoire locale </item>
    </list>
  </keywords>
</textClass>
```

FIGURE 4.7 – Capture d'écran : indexation matière RAMEAU des notices du dictionnaire.

L'indexation matière consiste en la description du contenu d'un document par le biais d'un vocabulaire normalisé et partagé au sein d'une communauté d'utilisateurs ou d'institutions. Cette indexation se fait donc par le biais d'une description en langage naturel, insérée au sein des métadonnées du document qui servent de base technique à son référencement¹⁶. Nous avons opté pour un référencement concernant le Dictionnaire et les thèmes de recherche abordés d'un point de vue global. Afin de s'inscrire à nouveau dans cette perspective de science ouverte, notre choix s'est porté sur une indexation utilisant le langage RAMEAU (Répertoire d'Autorité-Matière Encyclopédique et Alphabétique Unifié), maintenu par la BnF¹⁷, en raison de sa richesse et de son interopérabilité¹⁸. Bien que RAMEAU fasse l'objet d'une réforme dans sa structure, il nous semble qu'il s'agisse d'un moyen d'assurer la pérennité du référencement du Dictionnaire sur le moyen et long terme¹⁹. En somme, nous avons décidé de proposer un schéma stricte d'encodage

16. Étienne Cavalié, *L'indexation matière en transition : de la réforme de Rameau à l'indexation automatique*, ISSN : 0184-0886, Paris, France, 2019, p.9-12.

17. Voir <https://rameau.bnf.fr/>

18. Soulignons ici que dans la perspective de dépôt et archivage des notices sur la plateforme Nakala d'Huma-Num, le langage RAMEAU doit être traduit (« mappé ») en Dublin-Core adapté à Nakala

19. *Ibid.*, Voir chapitre 3 : « La réforme RAMEAU, principes, méthode, étapes ».

des métadonnées afin d’assurer d’une part l’insertion de chaque notice dans les principes directeurs de la science ouverte, et d’autre part l’automatisation de l’encodage des métadonnées (dans le <teiHeader>) sur l’ensemble du corpus.

4.1.3 Sources et références scientifiques au sein des notices

Un dernier point essentiel sur lequel nous nous proposons de revenir brièvement est la schématisation et l’encodage des références bibliographiques et des sources au sein des notices. En effet, rappelons que chaque notice se sépare en trois parties : son titre, sa définition et ses « références bibliographiques et sources historiques ». L’encodage et la schématisation de cette dernière partie a été un des enjeux clés de la formalisation de cet O.D.D. Effectivement, lors de notre encodage prospectif, nous avions proposé de séparer sémantiquement, dans la lignée de la tradition des études historiques françaises, les sources primaires des références bibliographiques. D’un point de vue structurel, il nous a fallu proposer une manière d’encoder à la fois ce « bloc des références et sources » dans sa globalité et chaque élément à part entière. Plusieurs possibilités d’encodage s’offraient à nous dans le cadre de l’encodage bibliographique d’après les *Guidelines* de la TEI²⁰ :

- Encoder les références bibliographiques dans une balise <biblStruct>, contenant les références bibliographiques structurées²¹.
- Utiliser la balise <biblFull> contenant l’ensemble des éléments TEI nécessaires à l’encodage bibliographique²².
- Avoir recours à la balise <bibl>, pouvant elle-même contenir un nombre indéfini de balise <bibl> pour chaque citation ou référence bibliographique²³.

Cette dernière option a été retenue lors de la schématisation en raison de sa permissivité. Dans la mesure où les références bibliographiques sont inégalement précises et ne respectent pas toutes les mêmes normes, la balise <bibl> nous permet d’encoder non seulement la structure, mais aussi son sens. Ainsi, une première balise <bibl> dont nous précisons l’attribut @type = ‘références’ contient pour chaque source ou référence bibliographique une autre balise <bibl>. Les sources sont quant à elles contenues d’une part dans les éléments <bibl> contenant une autre balise <idno type=’ArchivalIdentifier’> lorsque nous disposons de la cote archivistique, ou dans une simple balise <bibl type = ‘sources’>. Voici un exemple de la structuration des références scientifiques pour la notice

20. Voir *Guidelines P5*, chapitre 3.12.1 *Methods of Encoding Bibliographic References and Lists of References*. <https://tei-c.org/release/doc/tei-p5-doc/en/html/CO.html#COBITY>

21. Voir <https://tei-c.org/release/doc/tei-p5-doc/en/html/ref-biblStruct.html>

22. <https://tei-c.org/release/doc/tei-p5-doc/en/html/ref-biblFull.html>

23. <https://tei-c.org/release/doc/tei-p5-doc/en/html/ref-bibl.html>

amidon :

```

<bibl type="references">
  <bibl><idno type="ArchivalIdentifier"> AD Somme, 1C 2449 à 2451, élection d'Amiens ;
    </idno>
  </bibl>
  <bibl>
    <idno type="ArchivalIdentifier"> AD Rhône, 4C 607 à 614, élection de Villefranche ; </idno>
  </bibl>
  <bibl>
    <idno type="ArchivalIdentifier"> AN, D VI, dossier 17, procès-verbal du 9 décembre 1789 ;
    </idno>
  </bibl>
  <bibl type="sources"> Vues générales sur l'impôt des aides, les inconvénients de sa suppression et la possibilité de sa réforme, 1789 ; </bibl>
  <bibl>
    <idno type="ArchivalIdentifier"> AN, G2 125, registre ; </idno>
  </bibl>
  <bibl type="sources"> Arrêt du conseil d'Etat qui nomme le sieur Henri Clavel régisseur des droits compris dans la Régie générale, 15 septembre 1780 ; </bibl>
  <bibl type="sources"> Jacques Necker, Compte-rendu au Roi, Paris, Imprimerie royale, janvier 1781, p. 91 et 106 ; </bibl>
  <bibl type="sources"> Jean-Louis Lefebvre de Bellande, Traité général des droits d'aides, 2 vol., Paris, chez Pierre Prault, 1760 ; </bibl>
  <bibl type="sources"> Pierre Clément, Lettres, instructions et mémoires de Colbert, t. II, Paris, p. CIII ; </bibl>
  <bibl type="sources"> Jean-Louis Moreau de Beaumont, Mémoires concernant les droits &impositions en Europe, tome 3, Paris, Imprimerie royale, 1769, p. 277-472 ; </bibl>
  <bibl> Thomas Boullu, « La transaction... », Université de Strasbourg, novembre 2019 ; </bibl>
  <bibl> Aline Logette, « La Régie générale au temps de Necker et de ses successeurs, 1777-1786 », Revue historique de droit français et étranger, 1982, n°3, vol. 60, p. 415-445 ; </bibl>
  <bibl> Gustave Dupont-Ferrier, « Histoire et signification du mot « aides » dans les institutions financières de la France, spécialement aux XIVe et XVe siècles », Bibliothèque de l'Ecole des Chartes, 1928, t. 89, p. 53-69 ; </bibl>
  <bibl> Thierry Claeys, Les institutions financières en France au XVIIe siècle, Paris, SPM, t. 1, 2011, p. 286-287. </bibl>
</bibl>
```

FIGURE 4.8 – Capture d'écran : exemple d'encodage des sources et de la bibliographie de la balise « amidon ».

Notons que l'encodage des sources primaires en elles-mêmes a été un point complexe de notre travail. En effet, nous avons choisi d'encoder les sources dans un balise `<bibl>` car, dans le cadre des recommandations de la TEI, une balise `<bibl>` structurante ne peut contenir qu'un nombre relativement limité d'autres balises (dont notamment `<bibl>` et `<idno>`). La balise `<idno>` renseigne un identifiant dans un langage normalisé associé à une entité physique ou intellectuelle²⁴. Couramment utilisé pour encoder les identifiants pérennes de type DOI (*Digital Object Identifier*) ou VIAF, il nous a semblé qu'il s'agissait d'un compromis efficace puisque cette balise est acceptée dans la balise `<bibl>`, ne demandant donc pas d'adaptation et modifications profondes des règles de la TEI. D'un point de vue de la schématisation, nous avons eu recours aux possibilités offertes par le

24. <https://tei-c.org/release/doc/tei-p5-doc/en/html/ref-idno.html>

langage *Schematron* pour préciser le contexte d’utilisation des balises <bibl> (au niveau structurel ou sémantique), comme l’indique cet extrait de l’O.D.D :

```

<elementSpec ident="bibl" mode="change">
    <gloss>Référence bibliographique.</gloss>
    <desc> permet d’encoder les références bibliographiques plus ou moins
        structurées dans le bloc des références scientifiques (bibliographie et
        sources imprimées).</desc>
    <constraintSpec scheme="schematron" ident="bibl">
        <constraint>
            <s:rule context="//tei:bibl[references]">
                <s:assert test="//tei:bibl or tei:bibl[@type='source']"> Le
                    premier élément bibl rencontré doit contenir un élément bibl
                    (l’élément bibl englobant les références scientifiques doit
                    contenir un élément bibl). </s:assert>
            </s:rule>
            <s:rule context="//tei:bibl[references]/tei:bibl">
                <s:assert
                    test="//tei:idno[@type='ArchivalIdentifier'] or tei:title
                    or tei:author or tei:date or tei:pubPlace or tei:extent or tei:publisher or tei:bibl"
                    > L’élément bibl contenu dans un élément bibl doit soit
                    contenir un attribut @type dont la valeur est
                    'ArchivalIdentifier' ou un élément title, auteur, date,
                    pubPlace, extent ou publisher. </s:assert>
                </s:rule>
            </constraint>
        </constraintSpec>
    </elementSpec>
```

FIGURE 4.9 – Capture d’écran : extrait de l’O.D.D appliqué à la balise <bibl>

4.2 L’O.D.D du corpus : le *Dictionnaire comme une unité intellectuelle*

4.2.1 De la nécessité (technique) d’un deuxième O.D.D : les méthodes « agiles » en action

Avant de développer plus en détail les raisons qui nous ont poussées à proposer un second O.D.D ayant un paradigme de schématisation sensiblement différent, il nous faut expliquer rapidement le contexte de mise en oeuvre du *Dictionnaire numérique de la Ferme générale*. S’inscrivant dans le cadre des méthodes dites « agiles²⁵ », notre travail s’est structuré autour d’un développement incrémental et itératif. Ainsi, après avoir proposé à la fin du mois de mai 2022 un premier jet d’un prototype d’application web développé dans le langage Python grâce au framework Flask et avoir recueilli l’approbation de l’équipe de coordination technique puis scientifique, nous avons pu nous focaliser sur l’intégration de l’ensemble des notices de la lettre A (soit 29 notices). Or, à ce stade,

25. Les méthodes « agiles » préconisent en effet une planification adaptative et surtout la production rapide d’un livrable technique faisant l’objet d’une intégration continue, incrémentale et itérative des fonctionnalités. Sur ce sujet, on peut se référer au *Manifeste pour le développement agile de logiciels*, publié en 2001 aux Etats-Unis : <http://agilemanifesto.org/iso/fr manifesto.html>

une contrainte technique s'est imposée à nous : la nécessité d'automatiser la visualisation des notices. S'il était envisageable pour un échantillon de 3 à 4 notices de créer des *templates*²⁶ spécifiques, cette solution n'est nullement viable pour un dictionnaire proposant à terme plus de 300 notices. Ainsi, nous avons dû proposer un modèle d'automatisation de la visualisation, que nous décrivons dans la troisième partie de ce mémoire. Or, ce processus se fonde sur un paradigme de schématisation sensiblement différent qui implique la rédaction d'une second O.D.D.

En effet, ce nouveau cadre technique implique d'avoir à notre disposition un seul fichier TEI que nous segmentons automatiquement lors de la visualisation dans l'application web. Pour cela, notre second schéma s'est structuré autour de l'idée de créer un corpus de notices et non plus une simple collection de notices individuelles. Ce nouveau paradigme schématique présente donc le double avantage de répondre à un impératif technique et de modélisation. En effet, l'inclusion de l'ensemble des notices dans un groupe textuel général permet donc de développer un modèle conceptuel de dictionnaire considéré comme une oeuvre cohérente, organique et à part entière. En conséquence se pose la question de la coexistence de deux O.D.D et donc deux schémas d'encodage au sein d'un même projet. Ce changement de paradigme ne rend t-il caduque l'O.D.D appliqué aux notices individuelles ?

En somme, il apparaît clairement que notre choix de maintenir deux O.D.D au sein du projet répond à deux enjeux différents :

- **L'O.D.D des notices individuelles** est appliqué lors de l'encodage automatique des notices (voir chapitre suivant) et permet de contrôler la conformité et qualité de ce processus, que ce soit d'un point de vue structurel ou sémantique. De plus, cet O.D.D permet de garantir la pérennité des notices encodées, puisqu'à l'issue du projet, celles-ci seront déposées sur la plateforme de dépôt Nakala. Or, dans ce cadre, il est recommandé de déposer l'ensemble des documents afin de récupérer un DOI permettant de l'identifier de manière pérenne. Dès lors, l'O.D.D individuelle s'insère non seulement dans le *workflow* du traitement de la donnée comme un point essentiel du contrôle des données, mais aussi dans une perspective d'archivage sur le long terme.
- **L'O.D.D du corpus** répond quant à lui non seulement à un besoin technique circonscrit au cadre de l'application web, mais aussi à un modèle conceptuel organique.

26. Les *templates* sont des fichiers au format HTML qui servent de « patrons » pour proposer une visualisation des notices dans l'environnement web du Dictionnaire.

Afin de synthétiser notre démarche, le schéma suivant illustre l’interaction et la coexistence des deux O.D.D au sein du projet :

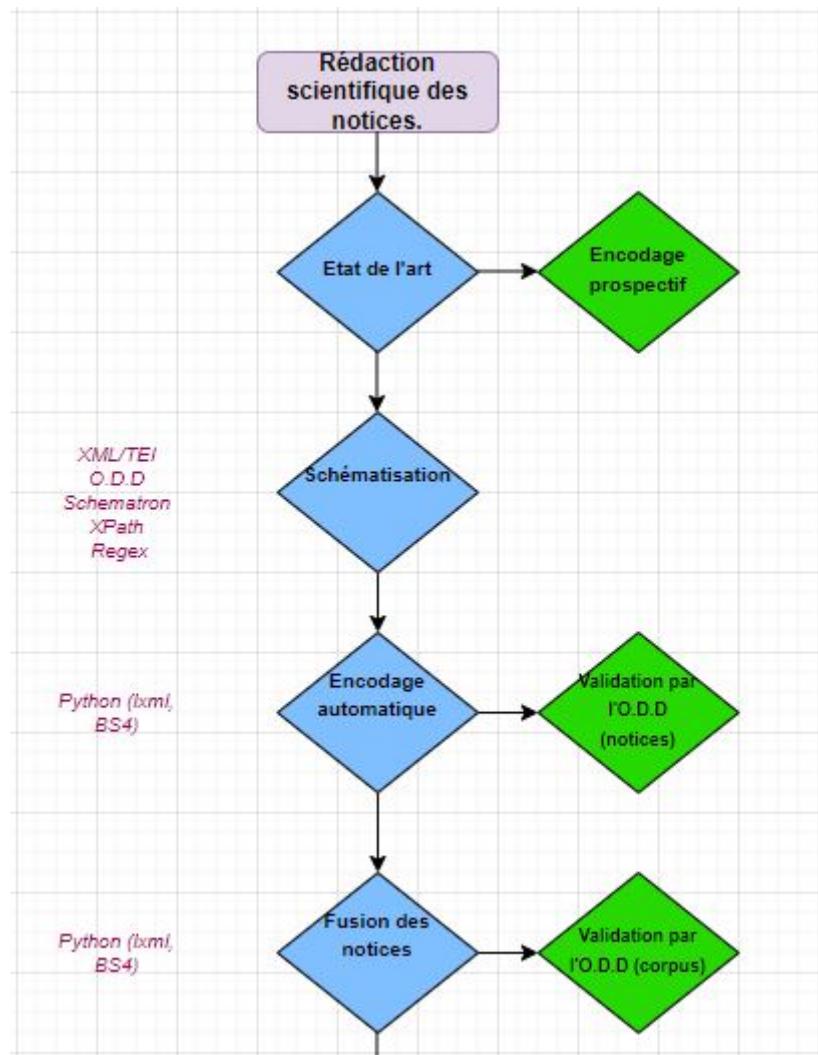


FIGURE 4.10 – Schéma : *workflow* et interaction des deux O.D.D

4.2.2 D’une collection de notice à une oeuvre organique : un autre paradigme schématique

D’un point de vue technique, ce nouveau paradigme schématique se traduit par l’incorporation de l’ensemble des notices dans une balise englobante, le `<teiCorpus>`²⁷. Cette balise est utilisée dans la grammaire de la TEI pour contenir des corpus de documents, eux-mêmes encodés dans des balises `<TEI>`. Ce corpus comporte ses propres métadonnées encodées dans une en-tête, le `<teiHeader>`. Par conséquent, l’arborescence du nouveau schéma prend la forme suivante :

27. Voir <https://tei-c.org/release/doc/tei-p5-doc/en/html/ref-teiCorpus.html>

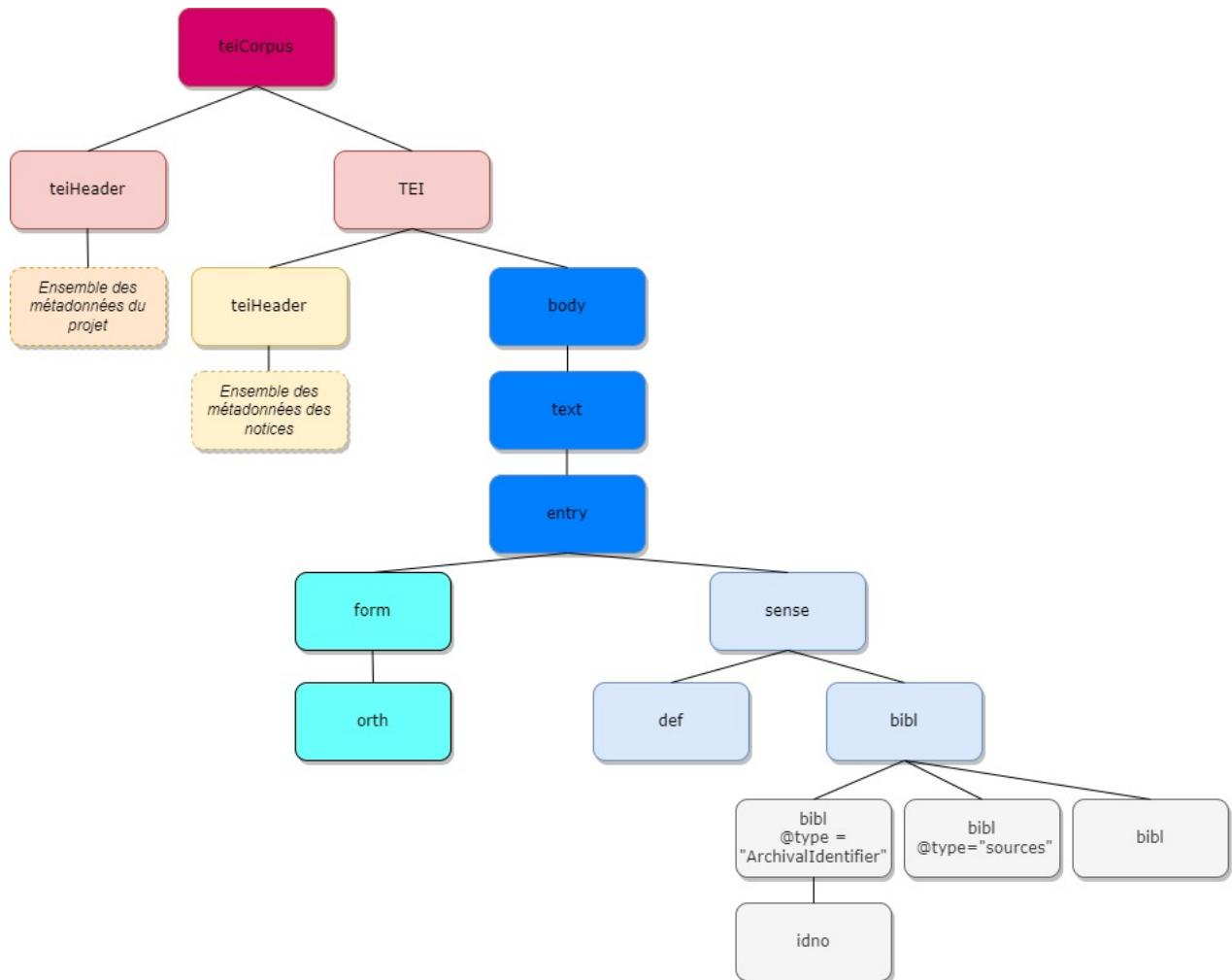


FIGURE 4.11 – Schéma : arborescence XML/TEI du corpus TEI pour l’application web.

Au sein de l’O.D.D en lui-même, nous avons donc dû ajouter la classe « `describedResource` » qui permet de déclarer l’utilisation d’un élément racine autre que l’élément `<TEI>`, en l’occurrence `<teiCorpus>`. A cet égard, il nous faut souligner une difficulté technique que nous avons rencontrée et qui s’avère être propre à cet enjeu de déclaration d’une nouvelle racine pour le document. Effectivement, rappelons que l’O.D.D est transformé grâce au logiciel Oxygen en fichier RNG qui sert par la suite à contrôler la conformité des fichiers TEI à notre schéma. Or, cette transformation automatique a systématiquement échoué dans la déclaration de la balise `<teiCorpus>` comme nouvelle racine. Il nous a donc fallu ajouter manuellement dans le fichier RNG transformé cette déclaration de nouvelle racine comme le montre l’extrait en question ci-dessous :

```

<start>
  <choice>
    <ref name="teiCorpus"/>
    <ref name="TEI"/>
  </choice>
</start>

```

FIGURE 4.12 – Capture d’écran : ajout du teiCorpus dans le fichier .rng à la main

Au niveau de cette seconde O.D.D d’autres modifications structurelles sont à mettre en évidence :

- Nous avons ajouté à ce nouvel élément racine `<teiCorpus>` un attribut obligatoire « `xmlns` » afin de déclarer l’utilisation de l’espace de nom TEI pour en assurer la conformité.
- Un second attribut « `version` » a été inséré au `<teiCorpus>` afin de pouvoir garder un historique des modifications du corpus²⁸.
- Aux éléments `<TEI>`, contenant chacun une notice, ont été ajouté un attribut « `n` » correspondant à son numéro, permettant la segmentation et visualisation des notices au sein de l’application web.

Dès lors, la présence de deux O.D.D au sein du même projet pose la question de leur coexistence et possible interaction. Une hypothèse que nous avons envisagée est le recours au principe de « chaînage d’O.D.D ».

4.2.3 L’hypothèse de l’*O.D.D chaining* : complémentarité ou coexistence des schémas d’encodage ?

Une première solution qui se présenta à nous fut celle du chaînage des O.D.D, ou *ODD chaining*, qui consiste en l’incorporation d’un O.D.D. dans un autre, de telle sorte qu’il est possible pour un document de réutiliser (et éventuellement de modifier) les déclarations et définitions d’un autre document. Cette technique de schématisation fut proposée par Lou Burnhard, co-fondateur de la TEI, dans le cadre d’expérimentations pour le Consortium TEI sur l’intégration et interaction de différentes parties d’O.D.D²⁹.

Expliquons brièvement les principes de l’*O.D.D chaining*. Par défaut, lorsqu’un O.D.D ne spécifie aucune source, les éléments sont collectés à partir de la version la

28. A cet égard, il est envisageable d’automatiser ce traçage des mise à jour par l’ajout d’un système d’incrémantation dans le script Python que nous décrivons dans le chapitre suivant.

29. L. Burnard, « ODD chaining for Beginners », <http://teic.github.io/TCW/howtoChain.html>

plus récente des *TEI Guidelines* mais ce comportement peut être modifié. Il est donc possible de changer cette référence et de mettre le chemin d'un autre O.D.D comme valeur de source. Selon ce principe l'O.D.D mère transmet ses propriétés à ses enfants. Ce sont donc tous les éléments communs qui doivent être écrits dans l'O.D.D mère, tandis que la structure, qui diverge, sera renseignée dans chaque O.D.D enfant. Sur ce sujet, notons qu'un chaînage d'O.D.D fut proposé dans le cadre du projet ANR TimeUs pour l'encodage et la structuration des minutes de procès aux Prud'hommes et des extraits de la presse, par V. Le Fournier, dont nous reprenons les explications.³⁰.

Nous avions donc envisagé le recours au chaînage d'O.D.D. Cependant, au-delà des enjeux techniques, ce processus soulève une question centrale qui est celle de la complémentarité ou de la coexistence des deux O.D.D. À bien des égards, nos deux O.D.D ne peuvent pas être chaînées puisqu'elles répondent à deux enjeux techniques et de conceptualisation sensiblement différents. En effet, d'une part l'O.D.D des notices individuelles répond, comme nous l'avons évoqué précédemment, à un besoin d'archivage pérenne alors que l'O.D.D du corpus répond avant tout à des problématiques techniques. De plus, l'O.D.D du <teiCorpus> ne pourrait pas être considéré à juste titre comme l'O.D.D mère puisqu'il n'a pas pour vocation d'interagir avec l'O.D.D des notices individuelles. En somme, l'O.D.D nous permet de proposer un contrôle de la conformité de la structure des notices, puis du dictionnaire aux besoins et enjeux du projet. Ce même dispositif de contrôle de la conformité doit de même s'appliquer aux entités nommées et à l'ontologie développée en accord avec l'équipe scientifique.

4.3 L'ontologie des entités nommées : comprendre l'emprise territoriale de la Ferme générale

4.3.1 Les enjeux scientifiques de l'encodage des entités nommées

Le dernier enjeu majeur de la rédaction de l'O.D.D est l'encodage des entités nommées en lien avec le projet scientifique du *Dictionnaire numérique de la Ferme générale*. Effectivement, le projet FermeGé entend comprendre l'emprise territoriale et spatiale de cette institution rationnelle mais discriminante qu'est la Ferme générale³¹. Ainsi, il a été nécessaire de formaliser une ontologie des éléments et entités à encoder répondant à cet

30. Les schémas de validation chaînés peuvent être trouvés à l'adresse suivante : <https://gitlab.inria.fr/almanach/time-us/schema-tei/-/tree/master/D%20-%20Sch%C3%A9ma%20de%20validation/D.2%20-%20DD> Victoria Le Fournier, *Étude de la structuration automatique et de l'édition d'un corpus hétérogène, l'exemple des sources du conseil des prud'hommes pour le textile du XIXe siècle*. Mémoire de master TNAH, Paris, 2019, p.69 - 73.

31. Pour rappel, voir la présentation scientifique du projet à l'adresse suivante : <https://dicofg.hypotheses.org/category/presentation-du-dictionnaire>

enjeux historiographique majeur. Pour ce faire, nous avons mis en place un instrument de travail en interne sous la forme d'un tableau partagé (framacalc) que nous reproduisons en annexe (voir Annexes - B.2).

Cette ontologie répond donc à la volonté de comprendre et encoder par l'établissement d'une typologie les différentes traductions spatiales de la Ferme générale³² :

- les « **pays primitifs** » correspondent aux territoires ayant conservé leur organisation propre ou leurs corps intermédiaires de représentation face à l'extension du pouvoir monarchique. Dans cette catégorie, se trouve les pays et territoires à diverses échelles qui revendiquent, tout particulièrement au XVIII^e siècle, ce statut « primitif » ou antérieur au royaume de France dans le cadre des théories provincialistes³³. Ainsi, sont encodés dans ce cas les duchés, principautés, seigneuries et pays d'états (par exemple : « Duché de Rethel, principauté de Sedan, seigneurie de Cheffes, duché d'Anjou, Gex, les Dombes, Labourd, Aunis, les Mauges »).
- les « **pays fiscaux** » viennent, dans l'ontologie du projet, en opposition aux « pays primitifs » puisqu'ils constituent l'ensemble des territoires structurés par les institutions financières royales, dont la Ferme générale. Les territoires où sont perçus la gabelle, dans toute leur diversité fiscale³⁴ sont ici encodés (« pays rédimés, pays d'aides, pays de petites gabelles, pays de gros, pays de quart-bouillon »).
- les « **provinces** » comprennent donc l'ensemble des provinces d'Ancien Régime. L'objectif scientifique est de comprendre la densité de cet ancrage de la Ferme générale à cette échelle particulière. Soulignons que les trois catégories décrites jusque ici font l'objet d'un encodage au sein de la balise <orgName> dans la mesure où nous estimons que le regard de l'équipe scientifique se pose dans ce cas sur les entités politiques qu'elles incarnent plutôt que sur des toponymes.
- les « **administrations et juridictions royales** » correspondent aux différents échelons de l'administration royale en terme financier ou juridique. Dans cette catégorie sont encodés les généralités, cours, commissions et élections mais aussi

32. Cette ontologie se base principalement sur les travaux en cours des notices du *Dictionnaire de la Ferme générale* et sur l'oeuvre de l'historienne Vida Azimi((V. Azimi, *Un modèle administratif de l'Ancien régime...*)).

33. Voir notamment M.-L. Legay, *Les États provinciaux dans la construction de l'État moderne aux XVII^e et XVIII^e siècles*, Genève, 2001 ; Bernard Barbiche, *Les institutions de la France sous la monarchie absolue, 1598-1789*, Paris, 2005

34. D. Dessert, *L'argent du sel : le sel de l'argent*, Paris, France, 2013.

les greniers car il s'agit d'institutions ayant une certaine emprise territoriale.

- les « **administrations et juridictions des fermes** » regroupent quant à elle les organisations propre à la Ferme générale. L'objectif est à la fois de distinguer l'action territoriale de la Ferme générale par rapport aux autres institutions royales, mais aussi de mettre en exergue la diversité des échelles mobilisées.
- les « **lieux d'habitations** » englobent l'ensemble des toponymes clairement identifiés concernés par l'implantation de la Ferme générale. Il s'agit ici principalement des villes ou ports. L'objectif sous-jacent est de comprendre l'interaction entre un espace spécifique ayant son propre fonctionnement et la Ferme générale.
- les « **lieux de contrôle** » encodent quant à eux l'ensemble des lieux ponctuels où s'effectue et s'ancre l'action de contrôle et de répression de la Ferme générale. Ceux-ci ont été différencié des lieux d'habitats dans la mesure où ils représentent des espaces de rupture géographique clairement identifiables et localisables.
- les « **espaces et territoires géographiques** » font l'objet d'un encodage dans une catégorie à part entière car il s'agit de zones plus diffuses et plus difficilement localisables. En effet, ce sont pour la plupart des espaces de contacts (au sens d'interface géographique) où la Ferme générale exerce son influence.

Ainsi, l'ontologie développée dans le cadre du projet est le fruit de choix scientifiques réalisés en amont du processus d'encodage automatique afin de pouvoir répondre à un enjeu historiographique précis. Il nous faut à présent mettre en évidence la manière dont nous avons schématisé, au sein de l'O.D.D cette ontologie.

4.3.2 Restreindre les valeurs des attributs par la grammaire O.D.D

Cette ontologie se traduit dans la schématisation par le recours à la grammaire de l'O.D.D qui permet de déclarer des listes fermées d'attributs et de valeurs d'attributs acceptées dans l'encodage. Soulignons que la TEI présente néanmoins une contrainte majeure en ce qui concerne la schématisation des types associés à une balise. En effet, il n'est possible de proposer que deux niveaux dans la typologie des éléments avec un attribut type et sous-type. Ceux-ci sont contenus dans la classe d'attributs « att.typed » que nous avons déclarée dans le préambule de notre O.D.D. Afin de restreindre la liste des valeurs autorisées, nous avons eu recours aux balises `<attList>` et `<valList>`, contenant

respectivement les balises <attDef> et <valItem> (liste des valeurs autorisées). Dans la mesure où cette ontologie est susceptible d'évoluer et de comprendre à terme plus de valeurs, nous avons décidé de laisser l'attribut « type='semi' » pour « semi-fermée » à la liste des valeurs³⁵. Voici les extraits de l'O.D.D pour les balises <orgName>, puis <placeName> en cohérence avec l'O.D.D.

```

<elementSpec ident="orgName" mode="change">
    <gloss>Nom de l'organisation</gloss>
    <desc> contient le nom d'une institution, administration, juridiction ou
          association de personnes. </desc>
    <attList>
        <attDef ident="type" mode="change" usage="opt">
            <desc> L'élément orgName doit contenir un attribut @type dont la
                  valeur est soit "province", soit "administration"</desc>
            <valList mode="add" type="semi">
                <valItem ident="province"/>
                <valItem ident="administration"/>
            </valList>
        </attDef>
        <attDef ident="subtype" mode="change" usage="opt">
            <desc>L'élément orgName doit contenir à la suite de l'attribut @type
                  un attribut @subtype dans la valeur est l'une des suivantes : </desc>
            <valList mode="add" type="semi">
                <valItem ident="administrations et juridictions royales"/>
                <valItem ident="administrations des fermes"/>
                <valItem ident="pays primitifs"/>
                <valItem ident="pays fiscaux"/>
            </valList>
        </attDef>
    </attList>
</elementSpec>
```

FIGURE 4.13 – Capture d'écran : extrait de l'ODD pour les organisations et institutions (orgName).

35. Ainsi, cet attribut permet de laisser la possibilité d'étendre la liste des valeurs mais n'autorise plus la modification de ses balises parents

```

<elementSpec ident="placeName" mode="change">
    <gloss>Nom d'un toponyme ou d'un territoire</gloss>
    <desc> contient le nom d'un toponyme, d'une aire géographique ou d'un
        territoire historique. </desc>
    <attList>
        <attDef ident="type" mode="change" usage="opt">
            <valList mode="add" type="semi">
                <valItem ident="toponymes"/>
                <valItem ident="zone"/>
            </valList>
        </attDef>
        <attDef ident="subtype" mode="change" usage="opt">
            <valList mode="add" type="semi">
                <valItem ident="lieux_habitation"/>
                <valItem ident="lieux_contrôle"/>
            </valList>
        </attDef>
    </attList>
</elementSpec>

```

FIGURE 4.14 – Capture d’écran : extrait de l’ODD pour les toponymes (placeName).

Dans un dernier temps, si cette ontologie et l’encodage des entités nommées se focalise sur les enjeux propres au *Dictionnaire numérique de la Ferme générale*, il nous faut souligner certaines limites et perspectives de développement futurs.

4.3.3 Les limites de l’approche géographique

Effectivement, notre ontologie ne prend pas en compte les personnages mentionnés au sein des notices et laisse la question des dates et de la chronologie au second plan. Il s’agit effectivement d’un choix de l’équipe scientifique de se focaliser sur l’espace plutôt que sur les personnes. Ce choix se justifie notamment par le fait que les travaux des historiens portant sur la Ferme générale comprennent déjà des listes importantes de noms des agents de la Ferme générale³⁶. Dès lors, si la balise <persName> est tout de même présente dans notre O.D.D, elle est utilisée pour encoder les noms de personnes uniquement dans les métadonnées des notices ou du corpus (en l’occurrence, les personnes responsables du projet).

Le cas des dates est quant à lui plus complexe. En effet, nous avons fait le choix de les encoder automatiquement dans la balise <date> consacrée à cet usage avec en ligne de vue l’objectif de proposer une chronologie en annexe du Dictionnaire. Cependant, l’équipe scientifique préfère fournir elle-même cette chronologie à partir d’un nombre restreint de dates significatives et commentées. Ainsi, un compromis fut trouvé avec la possibilité de fournir une liste exhaustive des dates présentes dans les notices que les historiens pourront à souhait réduire et commenter. Par cette approche quasi quantitative des mentions de dates et de périodes au sein des écrits scientifiques, il est probable que de nouvelles dates

36. *Ibid.*, voir annexes.

clés apparaissent et amènent hypothétiquement à repenser les évolutions chronologiques de la Ferme générale. En tout état de cause, d'un point de vue de la schématisation, nous avons du avoir recours au *Schematron* pour préciser la valeur des attributs des « when » ou « from » et « to » des balises <date>. Grâce à la déclaration d'une règle XPath nous avons d'abord indiqué que la balise <date> doit contenir ces attributs uniquement lorsqu'elle est enfant d'une balise <def> (en somme, lorsque les dates se trouvent dans la définition et non dans les métadonnées du <teiHeader> afin de ne pas fausser l'indexation). Voici l'extrait de l'O.D.D concerné :

```

<elementSpec ident="date" mode="change">
    <gloss>Date</gloss>
    <desc> contient la date d'un événement historique ou d'une période. </desc>
    <!-- Version simple et non restrictive qui n'impose directement la forme ISO 8601
        <attList>
            <attDef ident="when"/>
            <attDef ident="from"/>
            <attDef ident="to"/>
        </attList>-->
    <!-- Version plus complexe qui prend en compte le format de la date suivant la norme ISO 8601
        <constraintSpec scheme="schematron" ident="date">
            <constraint>
                <s:rule context="//tei:def//tei:date">
                    <s:assert test="@when or @from and @to"> L'élément date doit
                        contenir un attribut @when ou les attributs @from et @to.
                    </s:assert>
                </s:rule>
                <s:rule context="//tei:date[@when or @from or @to]">
                    <s:assert test="matches(., '^\\d{4}(-\\d\\d(-\\d\\d)?)?$')"> La
                        valeur de la date doit correspondre la norme ISO 8601 (soit
                        AAAA[-MM-DD] où les éléments entre crochets sont optionnels)
                    </s:assert>
                </s:rule>
            </constraint>
        </constraintSpec>
        <attList>
            <attDef ident="corresp" mode="delete"/>
            <attDef ident="facs" mode="delete"/>
            <attDef ident="type" mode="delete"/>
        </attList>
    </elementSpec>
```

FIGURE 4.15 – Capture d'écran : extrait de l'O.D.D concernant les dates.

En conclusion, la schématisation des règles d'encodage a été un point clé de notre travail dans la mesure où il s'agit à la fois du point d'entrée dans la chaîne de traitement, mais aussi un point de contrôle de la conformité et de la qualité des données. Ce duo d'O.D.D que nous avons mis en place entend donc répondre à deux impératifs distincts mais complémentaires, dans la perspective d'inscription du projet dans les principes de la science ouverte. L'encodage des entités nommées, quant à lui, est le fruit d'une élaboration en accord avec les choix et les enjeux scientifiques du projet FermeGé. Cette

phase de schématisation du projet a été réalisé en lien avec l'établissement d'un processus d'encodage automatique des notices qu'il nous faut à présent étudier.

Chapitre 5

Le processus d'automatisation de l'encodage : des notices au Dictionnaire numérique

Entre la schématisation d'une part et la mise à disposition des données d'autre part, la phase de traitement de la donnée et d'encodage automatique représente la clé de voûte de notre *workflow*. Afin de mettre en perspective ce travail de développement d'un encodage automatique, rappelons brièvement qu'un encodage manuel de l'ensemble des notices n'était nullement envisageable. Effectivement, ce corpus de 299 notices en avril 2022 est résolument évolutif puisqu'il est destiné à être complété jusqu'en 2024. L'enjeu principal de cette chaîne de traitement automatisée est donc de proposer un script capable d'encoder la structure des notices individuelles et les entités nommées que l'équipe scientifique désire voir indexées. Ensuite, afin de permettre l'intégration dans l'application web en cours de développement, cette collection de notices doit être rassemblée en un seul fichier, au sein d'une balise <teiCorpus>. Cet encodage doit ainsi intégrer un renvoi vers le schéma RNG des deux O.D.D. concernées et surtout respecter les règles formelles établies. Au-delà des nécessités de respecter les modèles conceptuels et logiques, un certain nombre de questions d'ordre technique se posent concernant les choix des modules adoptés. Ainsi, à travers ce chapitre, il nous faut analyser en quoi le processus d'automatisation de l'encodage implique un dialogue constant entre les choix techniques, les obligations des schémas et les perspectives de mise à disposition des données.

Au cours de ce chapitre nous proposons de mettre en perspective nos choix de développement et d'en montrer les avantages mais aussi leurs limites. Nous encourageons fortement nos lecteurs à se référer aux commentaires présents au fil du code des fichiers Python concernées dans les annexes (voir Annexes C.-5). En outre, pour mémoire, ce processus d'encodage s'inscrit dans la troisième phase clé de notre *workflow* que nous

illustrent dans le schéma suivant :

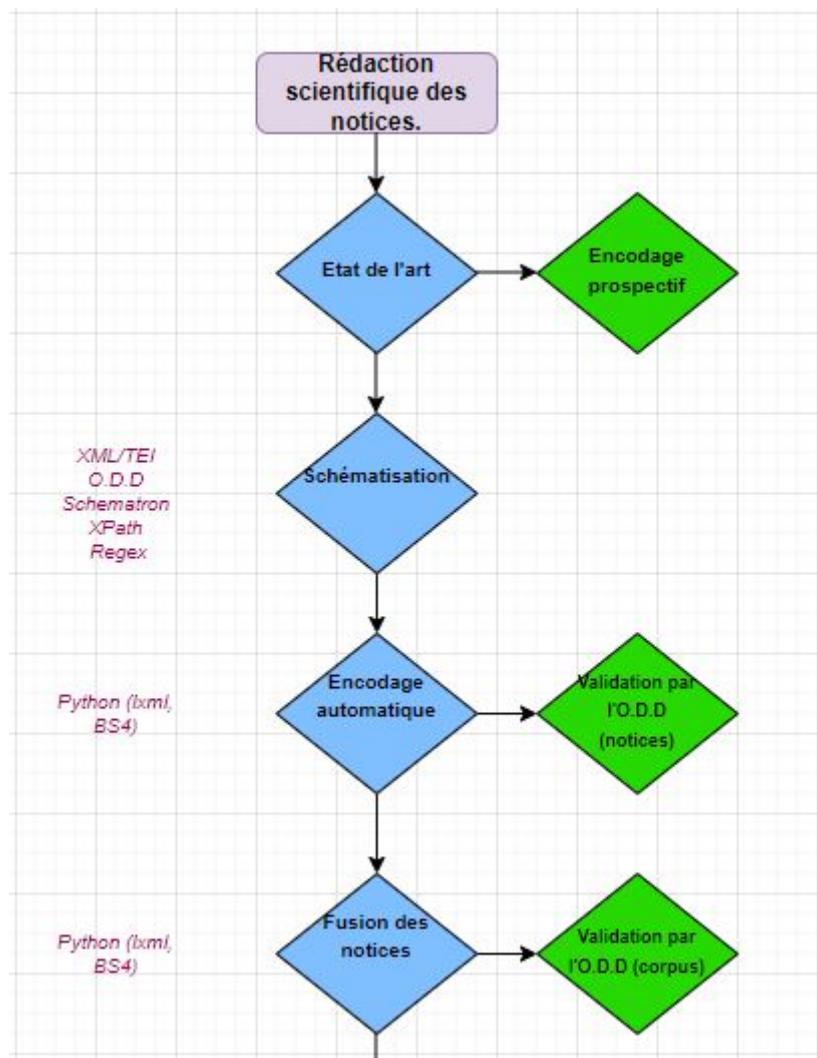


FIGURE 5.1 – *Workflow* : phase d'encodage automatisé des notices.

5.1 Segmentation et encodage de la structure : mise en application du schéma O.D.D

5.1.1 Le choix des armes : traiter des données XML avec Python

En premier lieu, nous avons utilisé le langage de programmation Python en raison de sa modularité et de sa flexibilité. Effectivement, nous sommes en mesure de proposer à la fois des scripts d'encodage automatique mais aussi par la suite un prototype d'application web. Cependant, l'une des difficultés de notre travail est de traiter des données XML de manière automatisée. Effectivement, l'XML/TEI est un métalangage décrivant sémantiquement des données et leur structure. Il ne s'agit donc pas à proprement parler d'un type de données natif implémenté en Python, à la différence des entiers (*integers*),

décimaux (*float*) ou des chaînes de caractères (*strings*)¹. C'est pourquoi nous avons besoin de « parser² » systématiquement des chaînes de caractères incluant la notation XML (par exemple la balise '') afin que celles-ci soient lues et comprises comme des balises XML. Dans cette optique, nous avons eu recours au module intitulé « BeautifulSoup 4 » (abrégé BS4) qui permet non seulement de « parser » les données mais aussi fournit un ensemble de fonctions de gestion, insertion et modification des balises ou attributs³. Le module BS4 s'adosse donc à un *parser* (« parseur ») XML qui est laissé au choix de l'utilisateur. La documentation du module présente à cet égard les avantages et limites des différents *parsers* disponibles, que nous reproduisons ci-dessous :

Parser	Typical usage	Advantages	Disadvantages
Python's html.parser	<code>BeautifulSoup(markup, "html.parser")</code>	<ul style="list-style-type: none"> Batteries included Decent speed Lenient (As of Python 3.2) 	<ul style="list-style-type: none"> Not as fast as lxml, less lenient than html5lib.
lxml's HTML parser	<code>BeautifulSoup(markup, "lxml")</code>	<ul style="list-style-type: none"> Very fast Lenient 	<ul style="list-style-type: none"> External dependency
lxml's XML parser	<code>BeautifulSoup(markup, "lxml-xml")</code> <code>BeautifulSoup(markup, "xml")</code>	<ul style="list-style-type: none"> Very fast The only currently supported XML parser 	<ul style="list-style-type: none"> External dependency
html5lib	<code>BeautifulSoup(markup, "html5lib")</code>	<ul style="list-style-type: none"> Extremely lenient Parses pages the same way a web browser does Creates valid HTML5 	<ul style="list-style-type: none"> Very slow External Python dependency

FIGURE 5.2 – Tableau de présentation des *parsers* en Python, extrait de la documentation BS4.

Notre choix s'est porté sur le *parser* intitulé *lxml's XML parser* en raison du besoin de gérer des données encodées en XML/TEI plutôt qu'en HTML (HyperText Markup Language).

5.1.2 Au début étaient les fonctions...

De prime abord, afin de réaliser ce script d'automatisation de pose des balises, nous avons dû mettre au point un ensemble de fonctions de manière séparée, qui sont ensuite appelées dans une fonction plus générale de structuration du document. Ainsi, chaque fonction à pour effet de créer un ou plusieurs arbres XML qui sont par la suite fusionnés

1. Emilien Schultz et Matthieu Bussonnier, *Python pour les SHS : introduction à la programmation pour le traitement de données*, Rennes, France, 2021, p.12-28

2. *Parser* signifie lire et segmenter une chaîne de caractère selon un critère spécifique, en l'occurrence la présence de balises

3. La documentation complète et structurée se trouve à l'adresse suivante : <https://www.crummy.com/software/BeautifulSoup/bs4/doc/>

pour donner naissance à un document XML/TEI valide et conforme à notre première O.D.D. Pour avoir un détail de ces fonctions, le lecteur peut se référer au code commenté et structuré en cinq parties consécutive présent en annexe (voir liens en Annexes A-7.3.3. Soulignons que dans cet ensemble de fonctions, nous en distinguons trois types :

1. Les **fonctions dites de « structuration »** qui permettent de créer des niveaux spécifiques de l'arborescence XML/TEI du document final. Ces différentes fonctions créent des morceaux d'arbres qui sont ensuite insérés ou mis bout à bout. La première de ces étapes est la fonction de création du « conteneur TEI », à savoir la balise racine qui accueille par la suite les autres balises (et ajoute le lien de renvoi vers l'O.D.D) que nous reproduisons en annexes (voir Annexes - Figure 8). Cette fonction *parse* donc une chaîne de caractères qui est transformée en arbre XML auquel nous ajoutons avec la méthode *.append()* un autre arbre créé indépendamment (en appelant la fonction « parsed tei header() »), puis une balise temporaire (« text blob ») qui sera par la suite modifiée. Ce même procédé est répété pour les balises de la structure, à savoir <body>, <entry>, <sense>, <orth>, <def> et <bibl>.
2. Les **fonctions « d'apport sémantique »** permettent quant à elles d'encoder dans des arbres XML restreints des passages du texte brut dans des balises sémantiques. Il s'agit donc principalement de l'ensemble des entités nommées (<orgName> et <placeName>) que nous analysons plus loin ou de certaines métadonnées. Soulignons que nous avons recours aux expressions régulières (abrégées *regex*) pour repérer et extraire les chaînes de caractères qui correspondent aux besoins de notre encodage. La fonction « add_title_to_metadata() » (voir Annexes - Figure 9) nous permet donc d'encoder automatiquement et ajouter dans les métadonnées de l'en-tête (<teiHeader>) le titre de la notice.
3. Les **fonctions dites de « correction » ou « nettoyage »** sont utilisées pour rendre valide l'encodage en fermant certaines balises orphelines en raison de la fusion des arbres, en supprimant certains caractères non voulus ou en normalisant les espaces blancs. Par exemple, la fonction « fix_idno_tag() » (voir Annexes - Figure 10) permet de corriger les balises fermantes <idno>.

5.1.3 Des « arbres » à la notice : processus de fusion et mise en cohérence de l'arborescence XML/TEI

Finalement, les fonctions centrales sont celles de « construction » de l'arborescence définitive des fichiers TEI de sortie. Ces fonctions appellent les différentes fonctions déve-

loppées au préalable qui sont sauvegardées dans des variables. Ces variables contiennent donc des arbres formant des fragments de l’arborescence que nous cherchons à créer. Il nous suffit dans un dernier temps d’ajouter et insérer ces fragments de manière progressive dans l’arborescence générale. Soulignons que c’est indéniablement dans cette perspective que les fonctions offertes par le module BS4 se révèlent les plus intéressantes. Effectivement, la méthode « `.insert(position, élément)` » permet d’insérer un fragment d’arbre dans l’arborescence générale avec une grande précision. De plus, la méthode « `.clear()` » a été utilisée pour normaliser l’imbrication des balises à l’échelle de la notice.

5.2 Encoder la bibliographie et les entités nommées : les limites de l’imbrication des balises

5.2.1 Un encodage de la structure par échelles

Afin de proposer un encodage de la bibliographie et des sources qui soit conforme à notre schéma d’encodage explicité dans les deux chapitres précédents, nous avons procédé à un encodage itératif et par échelles. En effet, comme nous l’avons mentionné précédemment, il nous a fallu repérer par le biais d’une expression régulière le « bloc des références scientifiques » (contenant la bibliographie et les sources). Une fois ce passage « capturé » dans le document texte original, nous l’insérons tout d’abord dans une balise générale `<bibl type='references'>`, correspondant à la balise `<bibl>` de niveau 1 dans notre schéma d’encodage. Une fois cette première opération achevée, nous « parsons » cette chaîne de caractères pour obtenir un nouveau fragment de l’arborescence. La capture d’écran en annexe (voir Annexes - Figure 11) illustre le fonctionnement de cette première fonction de structuration.

Dans un second temps, il nous faut encoder l’ensemble des références, ligne par ligne, dans une autre balise `<bibl>` qui sera par la suite traitée afin d’affiner l’encodage. L’objectif est tout d’abord d’encoder l’ensemble des citations bibliographiques et des sources de ce « bloc des références » dans une suite de balise `<bibl>` pour obtenir l’encodage structurel souhaité. Pour ce faire, nous avons tout d’abord recours à une autre expression régulière capturant les lignes de ce bloc de manière individuelle. Ces différents passages une fois capturés sont ajoutés à une liste sur laquelle nous itérons pour ajouter les balises `<bibl>` voulues. Indiquons que nous nous servons avant tout des marqueurs typographiques pour parvenir à notre fin. Effectivement, au sein du document texte original, les sources primaires et les références bibliographiques sont séparées par un point-virgule. C’est donc à partir de ce signe de ponctuation que nous effectuons la séparation des différentes chaînes de caractères, et par extension des différentes lignes. L’extrait de notre code correspondant à cette seconde phase d’encodage structurel des notices a été reproduit en annexe

(voir Annexes - Figure 12).

Ainsi, nous disposons dans ce cas d'un encodage de la structure générale de la bibliographie et des sources qui est conforme à notre encodage initial. Il nous faut à présent ajouter une couche sémantique correspondant aux enjeux du projet, à savoir l'indexation des références et, à terme, lier ces données avec des référentiels pérennes. En conséquence, nous devons compléter notre chaîne de traitement de la donnée par un processus de séparation de la bibliographie des sources.

5.2.2 Différencier la bibliographie des sources

En effet, la dernière étape de cette chaîne de traitement comprend donc l'ajout d'une couche sémantique dans l'encodage en permettant de différencier la bibliographie des sources. En premier lieu, les sources sont indéniablement l'élément le plus complexe à capturer et à encoder automatiquement. De plus, afin de gagner en granularité d'encodage, notre schéma O.D.D implique de différencier les sources primaires issues des fonds d'archives dépouillés des sources publiées. Ainsi, nous avons recours à deux expressions régulières répondant à ces enjeux techniques qui permettent d'encoder :

1. Les sources primaires issues des fonds d'archives comportant donc une cote qui nous sert d'élément de différenciation. Voici une liste des abréviations des fonds présents dans les notices :
 - **AM** pour archives municipales (suivi du nom de la commune concernée).
 - **AN** pour archives nationales.
 - **AD** pour archives départementales (suivi du numéro du département)
 - **BNF** pour Bibliothèque nationale de France
2. Les sources publiées tirées d'éditions scientifiques sont repérées et encodées grâce au type de document mentionné. La liste suivante recense cette typologie de documents que nous utilisons pour l'encodage :
 - **Edit** ou **Édit**, **Ordonnance**, **Règlement**, **Traité** et **Arrêt** ou **Arrest** pour les documents normatifs issus du pouvoir royal ou de sa justice déléguée.
 - **Carte** ou **Vue** pour les sources iconographiques ou à portée géographique.
 - **Almanach** et **Compte** pour les sources à caractère sériel.
3. Les **références bibliographiques** correspondent donc au autres chaînes de caractères qui n'entrent pas dans les critères ci-dessus et qui sont par conséquent encodés automatiquement dans une simple balise <bibl>.

Une fois ces expressions régulières formalisées, nous ajoutons à cette fonction les éléments ou attributs additionnels requis par l’O.D.D : soit la balise <idno type='ArchivalIdentifier'> pour les cotes archivistiques, soit l’attribut @type='sources' pour les sources éditées. La fonction associée à ce processus de traitement de la donnée se trouve en annexe (voir Annexes - Figure 13).

En guise de synthèse, le diagramme de flux ci-dessous entend schématiser le fonctionnement algorithmique du processus d’encodage des sources et références bibliographiques :

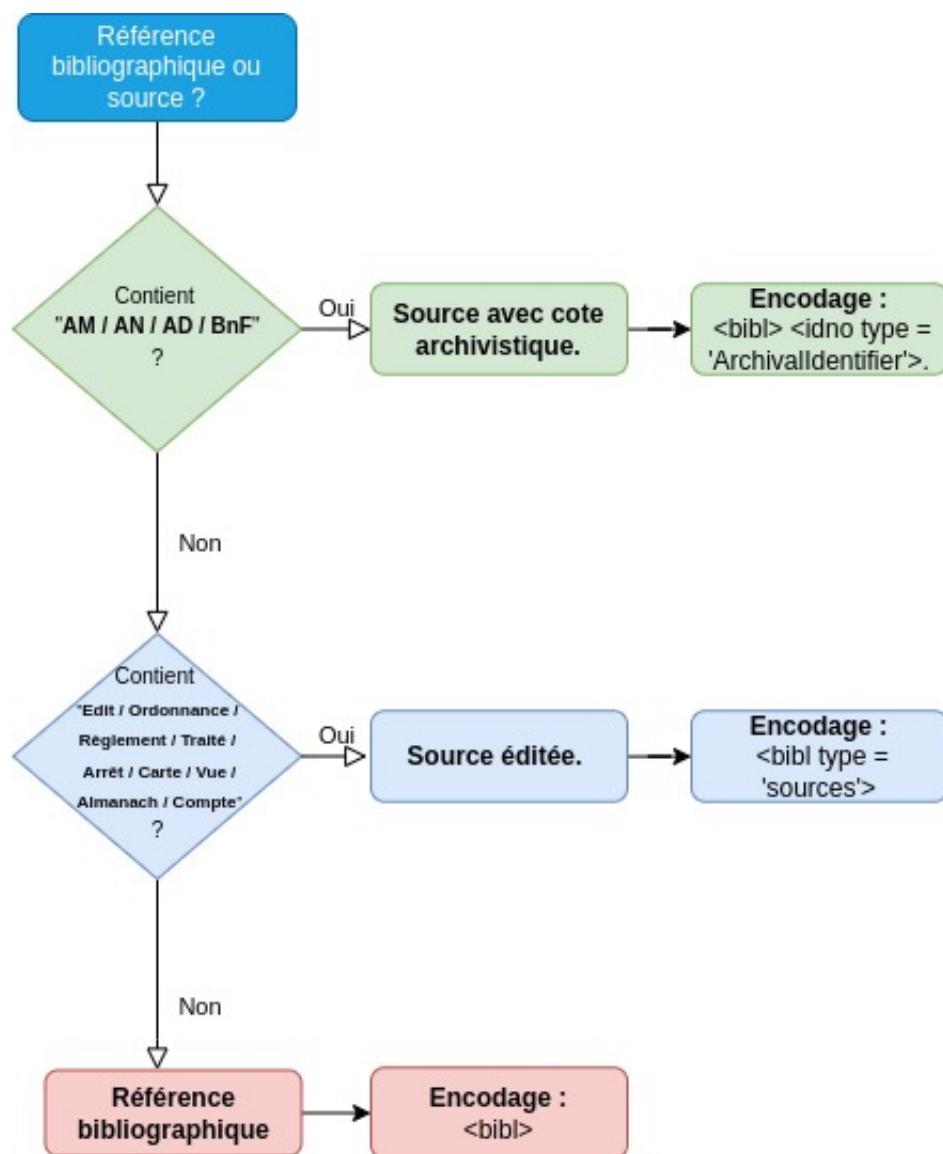


FIGURE 5.3 – Diagramme : fonctionnement algorithmique de l’encodage de la bibliographie et des sources.

Cet encodage automatique des références bibliographiques et des sources est donc assuré par un recours aux expressions régulières. Néanmoins, cette approche ne permet pas d’obtenir un encodage parfait et nécessite une part de corrections manuelles qu’il nous faut à présent décrire.

5.2.3 De nécessaires corrections : les limites de l'automatisation du processus

Afin d'évaluer la fiabilité de ce processus de traitement de la donnée et d'encodage des références scientifiques, nous avons procédé à un test sur un échantillon de 10 notices hétérogènes⁴, dont 2 comportant uniquement des renvois vers d'autres notices⁵. L'encodage s'est révélé satisfaisant et donc conforme à l'O.D.D dans 9 cas sur 10. Les erreurs générées venaient de l'ambiguïté de certaines publications qui étaient encodées comme des sources alors qu'il s'agissait de références bibliographiques. Par exemple, dans la notice « Acquit à caution », l'expression « *Dictionnaire des finances, 1727* » a été automatiquement encodé comme étant une référence bibliographique (au sens d'un travail scientifique contemporain) alors qu'il s'agit de toute évidence d'une source éditée. En conséquence, une phase de relecture et de correction manuelle est nécessaire afin d'assurer la conformité et la désambiguisation des données⁶

En tout état de cause, ce processus de traitement est satisfaisant mais ouvre certaines perspectives d'amélioration. En effet, il est nécessaire à terme de proposer un processus d'automatisation et de liens vers des identifiants pérennes concernant la bibliographie et les sources. Il nous semble effectivement envisageable d'ajouter une fonction de requête de l'API du *Worldcat* pour récupérer l'identifiant pérenne des références mentionnées ou obtenir leur DOI. Cette perspective s'inscrit dans les principes de la science ouverte et du web de données qui est le cœur de l'axe 4 du projet A.N.R FermeGé.

En somme, le processus d'encodage automatique concernant tout particulièrement la bibliographie et les sources nous a semblé particulièrement intéressant puisqu'il synthétise et met en perspective les enjeux techniques et scientifiques du projet.

4. Liste des notices de test : « Acquit à caution », « Adjudicataire », « Agenais », « Agent », « Agent général des fermes au Canada », « Aides », « Aigues-Mortes », « Allège », « Alsace » et « Alun »

5. En l'occurrence : « Agenais » et « Agent général des fermes au Canada »

6. Sur ce sujet, nous pouvons nous référer aux travaux de Maud Ehrman : Maud Ehrmann, *Les entités nommées, de la linguistique au TAL : statut théorique et méthodes de désambiguïsation*, These de doctorat, Paris 7, [s.l.], 1 janvier 2008. <http://www.theses.fr/2008PA070095..> Consulté le 14 avril 2022.

5.3 Des notices au *Dictionnaire* : une chaîne de transformation des documents

5.3.1 Appliquer le nouveau paradigme structurel : vers un processus d'encodage complémentaire

Dans un dernier temps, il nous reste à expliquer le procédé de fusion des notices individuelles au sein d'un corpus TEI afin de pouvoir satisfaire les besoins techniques de l'application web. Il s'agit donc dans cette dernière partie plus d'un changement d'échelle au sein de l'encodage qu'un second processus de traitement de la donnée. Comme nous l'avons précédemment mentionné, nous sommes confrontés à la nécessité de devoir développer un second script d'encodage afin de « fusionner » les notices individuelles et obtenir un corpus au sein d'un seul et même document, dans le cadre de la constitution de l'application web. Dès lors, le procédé utilisé est relativement similaire à la première phase de traitement et d'encodage des données. Il se déroule de la manière suivante :

1. Nous commençons par créer et « parser » l'**en-tête du nouveau corpus** à partir d'une chaîne de caractères qui renseigne l'ensemble des métadonnées à l'échelle du projet. Nous gardons la même approche pour la gestion des métadonnées que celle développée avec l'encodage des notices individuelles, à savoir une description des données administratives du projet⁷, un résumé des enjeux scientifiques et historiographiques de l'A.N.R⁸ et une indexation des informations techniques⁹.

2. Ensuite, nous avons recours à une seconde **fonction de « construction » de l'arborescence TEI** qui procède en quatre étapes :
 - une lecture et « fusion » de l'ensemble des notices individuelles enregistrées dans un répertoire dédié¹⁰. Nous avons inséré cette fonction en Annexes - Figure 14).
 - leur incorporation dans un fichier temporaire intitulé « corpus intermédiaire » (voir Annexes - Figure 15).
 - leur mise en conformité avec les attendus du schéma O.D.D associé¹¹, voir Annexes - Figure 16.

7. Éléments renseignés dans le <titleStmt>.

8. Rédigé en prose dans la balise <projectDesc>.

9. Disséminés dans les balises <encodingDesc> ou <appInfo>.

10. Nous avons ici recours au module *shutil* permettant de mettre bout à bout le contenu textuel d'un ensemble de fichiers présents dans le répertoire scripts/output et aux format XML (le « joker » * permet de signifier au script que nous voulons ouvrir et fusionner l'ensemble des fichiers portant pour extension « .xml ») La documentation du module *shutil* se trouve à l'adresse suivante : <https://docs.python.org/3/library/shutil.html>

11. Dans cette phase de traitement de la donnée, nous ajoutons tout d'abord l'espace de nom TEI à la nouvelle balise racine du document (le <teiCorpus>), puis nous ajoutons le nouveau <teiHeader> en première position après la balise <teiCorpus>

- Finalement, ce nouveau document conforme dans sa structure à l’O.D.D est enregistré et prend donc la forme du corpus final (voir Annexes - Figure 17).

5.3.2 Mise en conformité des données du Dictionnaire

Finalement, ce nouveau document conforme dans sa structure à l’O.D.D est « nettoyé et corrigé » par l’ajout de deux fonctions permettant d’ajouter des attributs « @n » aux balises <TEI> et <entry> et de supprimer les déclarations d’espaces de noms indésirables. Dans un dernier temps, la fonction en question permet d’enregistrer le résultat dans un nouveau fichier au format XML qui s’avère être la forme finale du corpus. L’ajout des attributs « @n » aux balises <TEI> et <entry> nous permet donc d’avoir un point de repère permettant la segmentation du corpus et la visualisation notices par notices au sein de l’application web. Cette décision fut prise à la suite d’un premier déploiement du prototype web, ce qui illustre à nouveau le travail itératif de développement du script et de l’application associée.

5.3.3 Vers l’application web : perspectives et limites du script d’encodage automatique

De manière rétrospective, si notre script d’encodage satisfait les besoins immédiats du projet, il nous faut mettre en évidence certaines perspectives d’amélioration et les limites de notre approche :

- En premier lieu, tout comme pour le processus d’encodage des notices individuelles, une phase de vérification et correction des éventuels défauts d’encodage est nécessaire. Par exemple, au sein des notices constituées uniquement d’un renvoi, nous devons nous assurer que la balise vide <def> (contenant la définition, ici absente, du lemme) soit bien fermée automatiquement. Lorsque des espaces blancs indésirables se sont malencontreusement glissés au sein de ce type de notice, il arrive dans de rares occasions que cette balise ne soit pas fermée.
- Le cas de ces notices constituées d’un simple renvoi vers une autre définition sont assez complexes à gérer en amont de l’intégration dans l’application web. Une possibilité d’amélioration serait d’ajouter automatiquement une balise <ref> (utilisée en TEI pour indiquer une référence vers un autre passage du texte) avec le contenu textuel du renvoi indiqué dans le titre de la notice.
- Le cas des graphiques, illustrations et documents d’archive numérisés présente de même un enjeu complexe d’automatisation de l’encodage. Effectivement, dans la mesure où nous travaillons à partir du texte brut extrait des notices, les images

n'apparaissent pas dans les données à la base de notre travail. Or, dans l'O.D.D nous avons pris la décision de baliser les documents iconographiques dans des balises `<figure>` et `<graphic>`. L'objectif final étant d'une part de pérenniser cette intégration des images et de permettre leur visualisation dans l'application web, par le biais d'une transformation XSL. Dans l'état actuel de la chaîne de traitement, nous ajoutons manuellement ce balisage des images au corpus final puisque celle-ci sont assez peu nombreuses¹². La difficulté principale étant l'impossibilité à ce jour de proposer une expression régulière capable de repérer les éléments typographiques ou sémantiques au sein du texte brut indiquant la présence initiale d'une image.

- Finalement, notre schéma d'encodage n'est pas en mesure de conserver pleinement la mise en page du document textuel initial puisque les différents paragraphes sont capturés et insérés au sein de la balise `<def>`. Afin de prévoir une visualisation et lecture plus agréable pour l'utilisateur de la version finale au sein de l'application web, il serait intéressant de proposer un moyen d'encoder automatiquement la présence de différents paragraphes dans des balises `<p>`.

En conclusion, le développement de ces deux scripts d'encodage, résolument complémentaires tout comme nos deux O.D.D, s'avère être la clé de voûte de notre travail, sous-tendant la conceptualisation ou formalisation du schéma d'une part, et la mise à disposition des données par le biais de l'application web d'autre part. Ce travail de développement a donc été pensé et réalisé de manière itérative et a permis de faire évoluer à la marge certains points du schéma d'encodage. Effectivement, les possibilités d'encodage automatique que nous avons mis en place reposent sur des motifs récurrents que nous repérons grâce à un ensemble d'expressions régulières. Dès lors, nous devons prévoir dans le schéma O.D.D un encodage suffisamment souple pour qu'une expression régulière soit utilisable dans le script. A cet égard, les entités nommées d'organisation ou de lieu ont été encodées dans une seule balise englobante `<orgName>`. En somme, ce travail de développement se veut être une réponse technique à une problématique scientifique initiale claire : comment assurer la pérennisation et l'encodage d'un corpus textuel brut relativement large ? Il nous faut finalement présenter le troisième livrable majeur de ce stage : le prototype d'application web du *Dictionnaire numérique de la Ferme générale* qui s'avère être à la fois une réponse au besoin de mise à disposition des données et une ouverture vers la science ouverte.

12. Pour le cas de l'échantillon traité des 29 notices de la lettre A, seules 5 illustrations sont à ce jour recensées. Dans la perspective de l'encodage de l'ensemble du corpus, la question de l'automatisation du processus doit néanmoins se poser.

*

De la mise en conformité des données à leur mise à disposition, il n'y a qu'un pas qu'il nous faut à présent franchir. Effectivement, dans un dernier temps, il nous semble nécessaire de mettre en évidence les mécanismes de diffusion et visualisation des données. Dans cette troisième et ultime partie, nous entendons illustrer la manière dont le développement d'un prototype d'application web s'affirme comme un moyen de répondre à un des enjeux principaux de l'ANR FermeGé : la diffusion des données. Au-delà des enjeux techniques de la mise en oeuvre du Dictionnaire dans sa version d'application web, il est nécessaire de souligner la manière dont ce travail doit s'inscrire, à terme, dans le contexte de la science ouverte et du web de données. Dès lors, nous proposons d'ouvrir notre réflexion à la possible insertion du *Dictionnaire de la Ferme générale* dans le champ plus vaste des données ouvertes et de leur pérennisation. Il est donc nécessaire de rappeler que la présente application sur laquelle s'appuie notre propos n'est qu'un prototype qui doit servir de base au développement du site web final.

En outre, la mise en oeuvre tant scientifique que technique du Dictionnaire s'insère dans l'axe 1 du projet ANR et communique de façon étroite avec l'axe 2 chargé de cartographier l'emprise territoriale de la Ferme générale. Notre réflexion doit prendre en compte la nécessité de prévoir l'intégration et le croisement des données géographiques, ayant leur modèle et format spécifique, avec les données textuelles au sein de l'application web.

*

Troisième partie

Le *Dictionnaire Numérique de la Ferme générale* : une esquisse de mise à disposition des données et ses enjeux

Chapitre 6

Le *Dictionnaire Numérique de la Ferme générale* : mettre à disposition les données

En premier lieu, comme nous l'avons précédemment souligné, la dernière partie de notre *workflow* se structure autour de l'enjeu de la mise à disposition des données au sein d'une application web. Si nous avons d'abord formalisé un modèle conceptuel et des règles d'encodage, puis dans un second temps, encodé automatiquement un échantillon de notices, il nous faut finalement permettre la visualisation de celui-ci au sein d'un prototype d'application web. Il s'agit dès lors du dernier livrable que nous présentons avant d'ouvrir la perspective du travail accompli pendant notre stage au champ de la science ouverte.

Avant toutes choses, nous entendons par « prototype d'application web » une ébauche de logiciel applicatif relevant des technologies, principes et protocoles développés dans le cadre du *World Wide Web* et permettant, à la différence des sites web statiques, une mise en application ou automatisation d'une tâche particulière, en l'occurrence la visualisation des notices. Une application web repose donc sur l'articulation de la technologie client-serveur et des requêtes du protocole HTTP¹. Ainsi, le navigateur web envoie au serveur une requête relative à un page web. Ce dernier émet un code de réponse permettant, lorsque la requête est valide, que le serveur est en ligne et que l'utilisateur possède tous les droits nécessaires, de visualiser la page demandée. Notre application web s'appuie donc sur ce fonctionnement pour permettre la matérialisation de son interface graphique. Dès lors, demandons-nous en quoi la phase de développement de l'application web permet de mettre en relation le processus de traitement de la donnée et son inscription dans les principes de la science ouverte.

En somme, il nous faut remettre dans son contexte le développement de l'application web et expliquer les choix techniques avant d'en présenter les principales fonctionnalités et

1. Les spécifications du protocole HTTP au coeur du fonctionnement de la technologie client-serveur sont maintenus à l'adresse suivante : <https://www.w3.org/Protocols/Specs.html>

les développements futurs à apporter. Le schéma ci-dessous entend rappeler brièvement l'état de notre *workflow* :

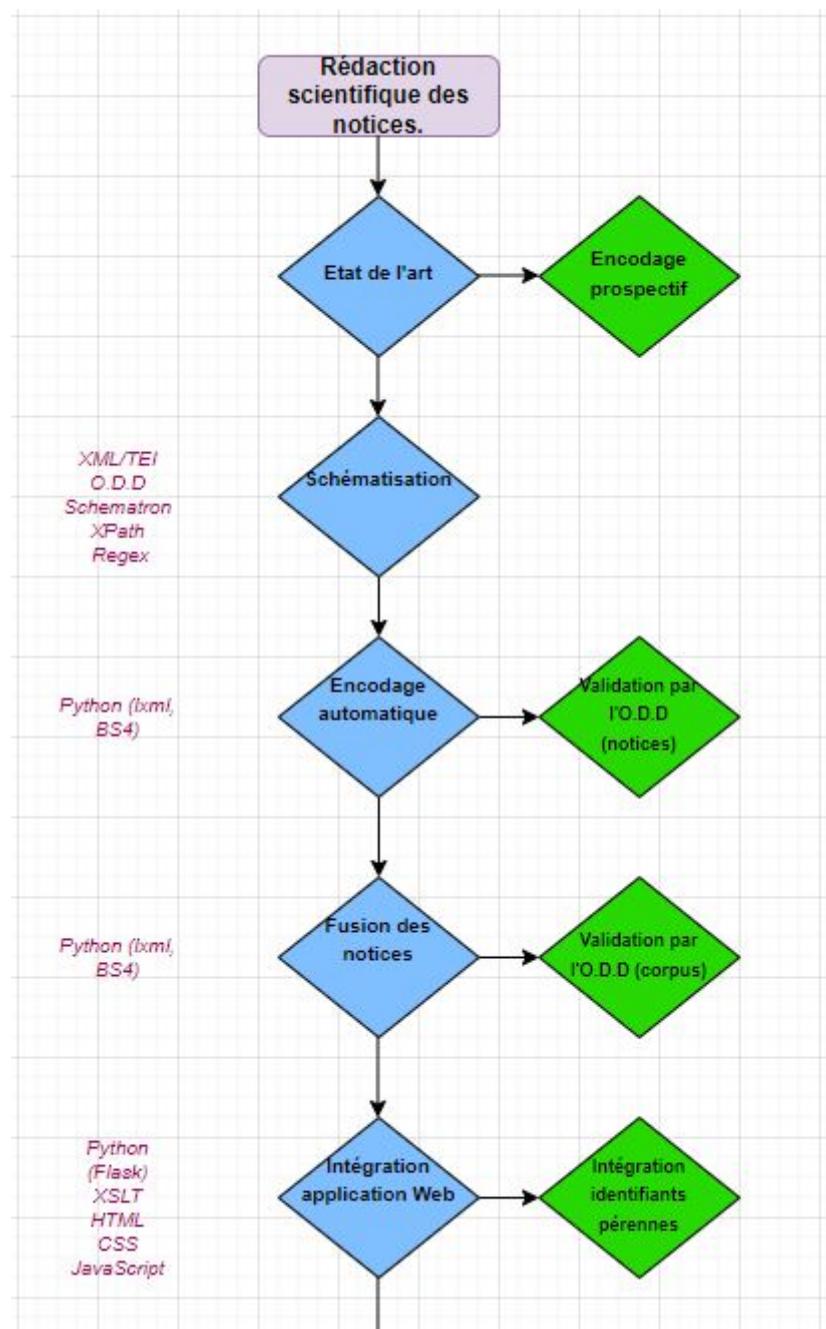


FIGURE 6.1 – *Workflow* : phase de développement de l’application web

6.1 L'application web : des choix techniques répondant aux besoins scientifiques

6.1.1 D'un *Content Managing System* à une application web : des choix en lien avec l'avancement du projet

Dans un premier temps, il convient de noter que le choix du développement d'une application web est le fruit d'une réflexion menée par l'équipe encadrante du projet dans la deuxième partie de notre stage. Effectivement, le *workflow* initial du projet prévoyait le développement d'une interface web par le biais du *Content Managing System* (CMS) Omeka ou Omeka S et le recours à un prestataire externe de développement web. Or, à partir de juin 2022, une réflexion s'est amorcée sur la possibilité de développer « en interne » une application web permettant non seulement la visualisation des notices, mais aussi la recherche et indexation des entités nommées. Le prototype que nous présentons se veut donc une ébauche de l'application web que pourrait devenir le *Dictionnaire numérique la Ferme générale*. Avant d'exposer et justifier les enjeux techniques qui nous ont poussés à adopter un développement applicatif en Python, rappelons quelques éléments clés sur Omeka et Omeka S. Omeka se présente comme « système de publication web spécialisé dans l'édition de collections muséales, de bibliothèques numériques et d'éditions savantes en ligne se situ[ant] à la croisée du système de gestion de contenus (CMS) de la gestion de collections patrimoniales ainsi que de l'édition d'archives numériques »². Dès lors, le recours à ce CMS aurait pu être un choix pertinent dans sa dimension d'édition et mise à disposition des données issues de la recherche en sciences humaines et sociales. Il est vrai que les notices dont nous disposons à ce stade du développement sont encodées dans un format, l'XML/TEI, propice aux éditions numériques. De plus, Omeka semble s'inscrire dans un écosystème de recherche et développement en humanités, puisque les créateurs et développeurs au sein du *Roy Rosenzweig Center for History and New Media* de l'Université George Mason (Virginie, États-Unis) sont aussi à l'origine du logiciel de gestion bibliographique libre Zotero. En outre, une communauté d'utilisateurs s'est structurée et semble faire corps autour du CMS, comme en témoigne les activités de l'Association des usagers francophones d'Omeka³. Omeka S se veut être la nouvelle version d'Omeka Classic permettant une plus grande inscription dans le web de données par l'implémentation de nouveaux *plug-ins*.

De même certains projets développés sous Omeka ou Omeka S présentent des similitudes plus ou moins grandes avec les données que nous mettons à disposition dans le

2. <https://omeka.fr/presentation-omeka>

3. Cet esprit communautaire semble se matérialiser notamment par les journées annuelles d'Omeka, les prochaines se tenant à Grenoble en novembre 2022 : <https://omeka.fr/journees-omeka-2022---grenoble-les-24-et-25-novembre>

Dictionnaire. A cet égard, nous pouvons citer les projets suivants :

- « Ontologie du christianisme médiéval en images⁴ », qui se présente comme une encyclopédie visuelle de la pensée chrétienne dans l’Occident médiéval. Ce site web met à disposition des chercheurs une synthèse encyclopédique constituée de courtes notices, appuyée par de nombreux documents visuels, concernant le christianisme médiéval. Ce site se rapproche de notre projet par l’aspect encyclopédique et les liens tissés entre les différentes notices.
- « Corpus et ressources archéologiques⁵ », est un site Omeka développé au sein du centre Camille Jullian (INIST) qui permet la visualisation d’une collection de notices ou fiches descriptives d’objets archéologiques. Ce projet est particulièrement intéressant pour nous dans le cadre de la structuration des données et l’ouverture des métadonnées au web de données.

Cependant, Omeka et Omeka S présentent un ensemble de limites qu’il nous faut expliquer, justifiant notre choix de procéder à un développement web en Python depuis la base jusqu’aux fonctionnalités les plus avancées. Comme le soulignent les historiens Cécile Boulaire et Roméo Carabelli dans leur chapitre consacré à Omeka tiré de l’ouvrage collectif *Expérimenter en humanités numériques*⁶, la rigidité d’Omeka présente des inconvénients multiples. Effectivement, la nécessité de renseigner les champs de métadonnées en *Dublin Core* ou *Dublin Core extended* et l’obligation d’adapter les champs renseignés à la structure d’Omeka s’avèrent être des contraintes. De même, le fonctionnement et la gestion des données au sein d’un CMS comme Omeka n’est que peu compatible avec le projet FermeGé en raison des enjeux suivants :

- Omeka est avant tout pensé pour la mise à disposition de données patrimoniales dans les domaines des musées, des bibliothèques et des instituts de recherche. Or, nos données sont très majoritairement textuelles et produites par le travail des chercheurs. Dès lors, en terme de visualisation et d’exposition virtuelle des données, les avantages d’Omeka pour notre projet sont assez faibles.
- De plus, malgré la diversité de ses *plug-ins* additionnels, Omeka et Omeka S ne disposent que d’outils embryonnaires pour l’indexation et la gestion des entités

4. <https://omci.inha.fr/s/ocmi/page/accueil>

5. <https://ccj-corea.cnrs.fr/>

6. Cécile Boulaire et Romeo Carabelli, « Chapitre 7. Du digital naïve au bricoleur numérique : les images et le logiciel Omeka », dans *Expérimenter les humanités numériques : Des outils individuels aux projets collectifs*, éd. É. Cavalié, F. Clavert, O. Legendre, et al., Montréal, 2018 (Parcours numérique). <http://books.openedition.org/pum/11115..> Consulté le 1 août 2022, p. 81-103. ISBN : 979-10-365-0173-9.

nommées. Or, dans le cadre du projet, cette indexation fera l'objet d'un approfondissement et développement par la suite. De même, la gestion des données structurées et encodées en XML semble assez complexe au sein de ce CMS.

- En dépit de la simplicité d'utilisation d'Omeka, nous préférons avoir un contrôle total sur le code produit au sein du projet, et pouvoir affiner ou structurer sans contraintes l'architecture et la mise en page du site en construction. Effectivement, puisque cette application n'est qu'un prototype, nous voulons garder la possibilité de développer nous même de futures fonctionnalités, sans dépendre du travail effectué par une autre équipe de développement.
- Finalement, en développant une application en Python, nous pouvons non seulement mettre à disposition notre code, et ainsi répondre aux impératifs de science ouverte de l'ANR, mais aussi nous inscrire dans une communauté internationale de développeurs utilisant l'un des langages informatiques les plus populaires : Python.

6.1.2 L'architecture : Flask comme *framework* de développement web

De prime abord, à partir de cette sous-partie nous fondons notre propos sur le prototype d'application web. Nous invitons le lecteur à se référer au guide d'installation et de lancement du prototype que nous avons rédigé en annexe D (voir Annexes - D.7).

Ainsi, en accord avec l'équipe scientifique et technique, nous avons développé un premier prototype d'application web en langage Python, utilisant le *framework* ou *micro-framework* Flask, ainsi qu'un ensemble de modules adaptés au développement web (*Flask-SqlAlchemy*, *Login-Manager*) et à la gestion des données XML (*Lxml* et *Beautiful Soup*⁴). Le module Python nommé Flask est un *framework* de développement applicatif web, à savoir un cadre de développement fournissant un ensemble de fonctions et méthodes permettant la génération automatisée de pages web, l'intégration d'une base de données ou le développement des moteurs de recherche et d'indexation⁷. Il est de coutume d'accorder à ce *framework* le terme de « micro », soulignant non seulement sa simplicité et sa flexibilité. En effet, par défaut Flask laisse au développeur le choix du type de base de données à utiliser et propose un très vaste ensemble de modules à ajouter afin de permettre par exemple de la recherche avancée (module *Whoosh*), de l'indexation approfondie (*Elastic-*

⁷. La documentation officielle se trouve à l'adresse suivante : <https://flask.palletsprojects.com/en/2.1.x/foreword/>

Search) ou même de la cartographie (modules dérivés de *Folium*). Ainsi, le problème de rigidité que pouvait présenter le CMS Omeka est ici résolu par l'adoption du *framework* en question.

De plus, Flask s'inscrit dans une communauté de développeurs et d'utilisateurs large et diverse au niveau mondial, permettant ainsi un dynamisme dans le développement de nouveaux modules et la gestion des possibles problèmes techniques. A cet égard, une étude datant de 2018 indiquait que Flask était le *framework* de développement web le plus populaire au monde avec 47 % du total des développeurs web en Python l'utilisant⁸. Ainsi, nous inscrivons notre travail dans un ensemble de bonnes pratiques et de recommandations à l'échelle la plus large possible, ce qui entre en résonance avec les impératifs d'ouverture des données au sein de l'ANR FermeGé.

D'un point de vue technique, Flask apporte certaines spécificités de développement et de fonctionnements qu'il nous faut brièvement expliciter⁹ :

- Flask permet de générer des pages web à partir de « routes », c'est à dire des chemins dans l'arborescence du site afin d'accéder à la page en question. Cela est rendu possible par un *decorator* modifiant le comportement des fonctions Python traditionnelles. Par exemple, la notice « aides » au sein de l'application web se trouve, à partir de la racine, dans le sous-domaine « notices » et dans son propre sous-domaine « 6 » (numéro de la notice en question), soit la route suivante : « /notices/6 ».
- La visualisation des pages web fonctionne par l'appel de *template* au format HTML, prenant la forme de « modèles structurants » destinés à accueillir les données des notices. Flask permet d'automatiser la création des pages web par la répétition automatique d'un *template* ou l'insertion d'un *template* « corps » dans un *template* général « conteneur » sur le modèle suivant :

8. L'étude en question se trouve à l'adresse suivante : <https://www.jetbrains.com/research/python-developers-survey-2018/>. Gardons néanmoins une grande prudence vis-à-vis de cette étude car la méthode d'enquête de terrain ou les critères du panel ne sont pas clairement renseignés. De plus, elle est partiellement conduite par une société privée *Jet Brains*, leader du marché dans les environnements de développement en Python (voir l'IDE *PyCharm*)

9. Notre travail s'est basé sur l'ouvrage de référence de Miguel Grinberg : Miguel Grinberg, *Flask Web Development : Developing Web Applications with Python*, Sebastopol, 2018.

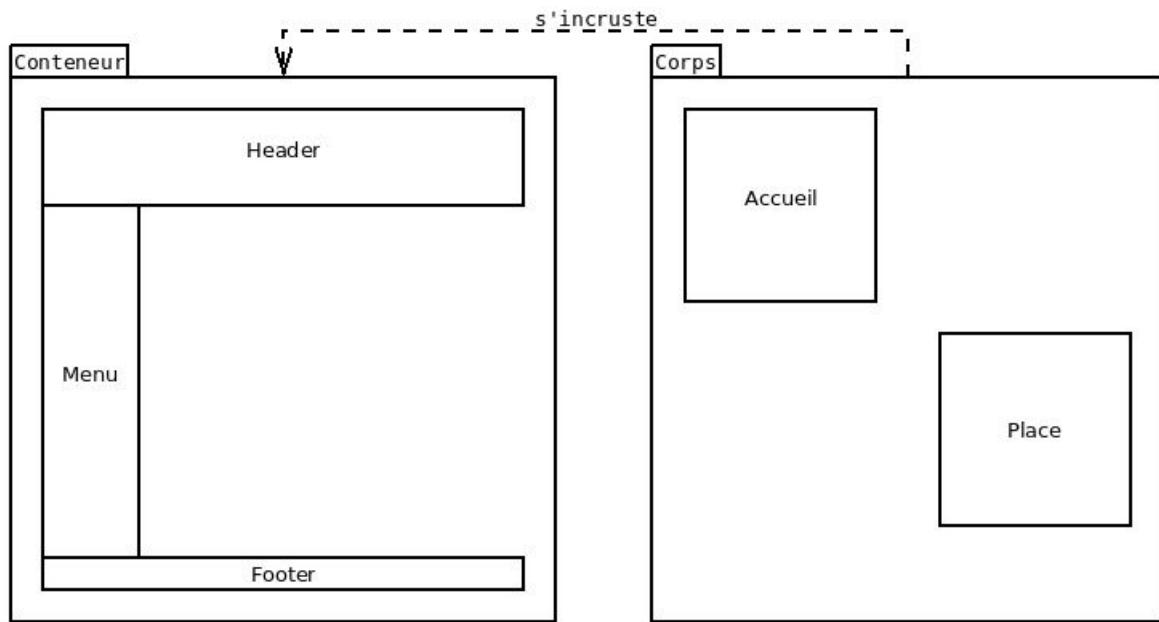


FIGURE 6.2 – Fonctionnement des *templates* avec Flask.

En somme, un développement d'une application web en Python avec le *framework* Flask présente l'avantage d'une très grande flexibilité de développement, nous permettant d'anticiper et suivre les évolutions techniques et les attentes de l'équipe technique et scientifique du projet. Effectivement, le caractère évolutif du projet nous pousse à nous inscrire dans une communauté de développement technique qui se veut la plus large possible. Ce choix d'un cadre de développement technique s'associe à un second *framework* assurant le développement des fonctionnalités visuelles et graphiques.

6.1.3 L'agencement graphique : Bootstrap comme *framework* graphique

Ainsi, au sein de notre application, le développement dit *backend*, relevant du développement technique et du fonctionnement « à l'arrière plan » est assuré par le *framework* Flask. L'aspect *frontend*, à savoir l'aspect graphique et visuel de l'application, est assuré par l'utilisation du *framework* nommé *Bootstrap 5.0*. Il s'agit d'une collection d'outils contenant des codes HTML pour l'architecture et la structuration des informations sur les pages web, ainsi qu'une feuille CSS¹⁰ utilisée pour la création du design et des extensions JavaScript¹¹ afin d'améliorer l'interactivité du site. On parle donc d'une « librairie » *Bootstrap* qui permet de structurer et rendre clair l'aspect visuel d'une page web. Cette

10. *Cascading Style Sheets* ou « Feuilles de Styles en Cascades » est un langage informatique à part entière qui permet de décrire par le biais de classe la forme que doit prendre un élément de l'architecture HTML d'une page web

11. Langage de programmation orienté web qui permet la modification et l'animation de pages web statiques.

véritable boîte à outil du développement *frontend* permet donc de générer un ensemble de boutons, de barres de recherche, de blocs de navigation et autres éléments de mise en page¹².

Bien que flexible et permissif, ce cadre de développement visuel présente un certain nombre de limites auxquelles nous avons du pallier. En effet, en terme de gestion des couleurs et de développement d'une identité visuelle propre, *Bootstrap* est relativement limitée. Il est effectivement possible de personnaliser et de déclarer certaines classes de couleurs en modifiant directement la feuille CSS initiale. Cependant, nous avons préféré dans notre cas insérer directement dans les attributs des éléments concernés des déclarations de classes CSS afin d'obtenir un résultat ponctuel précis. Par exemple, la chronologie développée dans l'onglet éponyme est le résultat d'une modification directe du comportement des éléments HMTL (onglet « Index et chronologie > chronologie »). De même, les couleurs employées dans l'application ont été déclarées directement dans la syntaxe CSS. A cet égard, la capture d'écran présente en Annexes E - Figure 18) illustre quelques choix graphiques effectués :

- Le menu de navigation reprend la couleur rouge tirée de l'extrait du plan des barrières d'octroi dans Paris présent sur cette page d'accueil.
- Le logo¹³ présent en haut à gauche de l'illustration reprend cette même couleur pour les lettres "FERME". La couleur rose clair est tirée d'un second document iconographique, que nous présentons plus bas¹⁴, dont les visages des personnages sont de cette même couleur.
- Le fond vert clair fait écho à la couleur des rues sur ce plan.

Ainsi, nous avons dans un premier temps étudié le contexte et les outils de développement s'offrant à nous afin de répondre aux enjeux scientifiques et techniques du projet. Le choix d'un *framework* développé en Python s'explique donc avant tout par la volonté de garder la main sur les possibles développements et demandes techniques à venir au cours du projet. Il nous faut donc présenter la forme concrète que prend cette application web.

12. La documentation officielle de Bootstrap se trouve à l'adresse suivante : <https://getbootstrap.com/>

13. Réalisé par Florence Perret, ingénierie en humanités numériques à la MESHS

14. Voir le tableau anonyme du XVIII^e siècle intitulé *Le Grenier à sel*

6.2 Tour d'horizon et état de l'art du *Dictionnaire Numérique de la Ferme générale*

A présent, il nous faut mettre en perspective les réalisations techniques et la forme que prend ce prototype de *Dictionnaire numérique de la Ferme générale*. Nous encourageons ainsi le lecteur soit à explorer le prototype par lui même en « clonant » le dépôt Gitlab correspondant et en suivant l'exécution des commandes décrites en annexe D (guide d'installation et utilisation - Annexes D .7). Au delà de la présentation des spécifications techniques, il nous faut amorcer une réflexion critique sur la forme et la place du Dictionnaire au sein du projet.

6.2.1 Visualiser les notices et naviguer dans le Dictionnaire

Dans un premier temps, le prototype d'application web répond à l'enjeu et à la demande principale du projet, à savoir permettre la visualisation des notices scientifiques concernant la Ferme générale. Il est possible d'accéder à la liste des notices depuis la page d'accueil en cliquant sur le menu déroulant à droite de la page intitulée « Liste des notices publiées ». Un index plus approfondi des notices comportant certaines métadonnées (auteur de la notice, identifiant et titre) se trouve dans l'onglet « Découvrir le projet » et « Axe 1 : Dictionnaire numérique de la Ferme générale ».

Dans le cadre de ce présent travail, nous avons intégré un échantillon de 29 notices correspondant à la lettre A. Ce choix, bien qu'arbitraire, permet de montrer la diversité des notices tant dans leur fond scientifique que dans leur forme. Effectivement, nous disposons à la fois de notices relevant de l'histoire économique et fiscale (« Aides »), des institutions et agents de la Ferme générale (« Agent »), des provinces (« Alsace », « Artois ») ou même de l'histoire matérielle (« Allège »). De plus, certaines notices comportent des illustrations (cliché des plomb de scellé tiré de la notice « Adjudicataire ») et documents de travail (graphiques et schémas produits par les historiens, comme au sein de la notice « Aides »). L'insertion des reproductions de ces divers documents pose un ensemble de problèmes d'ordre juridique que nous exposons et étudions plus loin. De même, soulignons qu'au sein des notices se trouvent deux types de liens cliquables, que nous avons mis en évidence avec deux couleurs différentes. Les liens en rouge (reprenant la charte graphique précédemment mise en évidence) permettront d'accéder aux autres notices liées à la notice en cours de consultation. En raison du caractère très partiel à cette date des notices, ces liens ne sont pas encore effectifs et demanderont un travail d'harmo-nisation par la suite. En outre, au bas des pages se trouvent une division spécifique de la page contenant d'une part les sources archivistiques et imprimées mobilisées, et d'autre

part la bibliographie scientifique mobilisée par les historiens. Cette mise en page répond donc au besoin de segmentation sémantique des informations établi précédemment dans le modèle conceptuel. Nous avons finalement automatisé la création d'un bloc de citations des notices qui permettra de récupérer automatiquement les éléments bibliographiques nécessaires à la citation du Dictionnaire, ainsi qu'une liste des notices liées qui pourraient à terme se présenter sous forme d'un graphe.

Cette visualisation des pages est le résultat d'une combinaison de différentes solutions techniques que nous avons structurée au sein de cette application web. Nous ne commenterons pas en détail ici notre code Python puisque ce travail relève de la documentation technique qui est disponible au sein du dépôt Gitlab. Nous proposons cependant de résumer le fonctionnement de l'application de visualisation des notices par ce schéma :

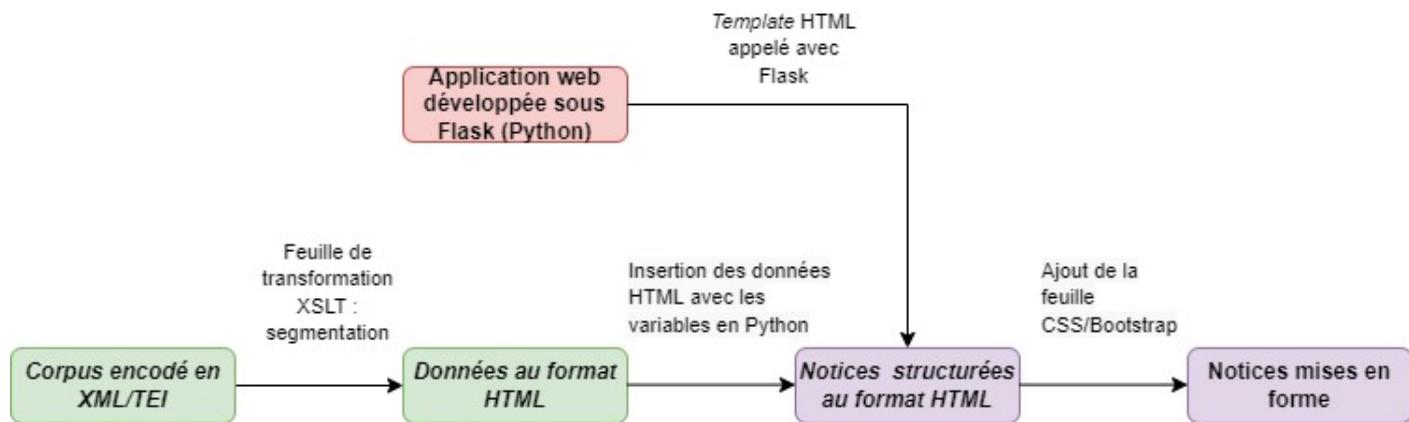


FIGURE 6.3 – Schéma fonctionnement de l'application web Flask.

6.2.2 Rechercher, requêter et explorer les notices ou index

Une seconde fonctionnalité majeure du *Dictionnaire numérique de la Ferme générale* est la possibilité de requêter et rechercher au sein du corpus un ensemble d'informations incluant non seulement les notices déjà évoquées, mais aussi les différentes entités nommées indexées en accord avec l'ontologie développée par l'équipe scientifique. Dans un premier temps, nous avons développé un moteur de recherche simple permettant de requêter les titres, identifiants et auteurs des notices. Pour ce faire, nous avons alimenté une base de données relationnelle par le biais d'un script récupérant un ensemble d'éléments XML/TEI au sein du corpus général¹⁵. Ensuite, par la déclaration de classes au sein d'un système de « mapping objet-relationnel (ORM) », nous pouvons grâce au script Python mettre en oeuvre un moteur de recherche, transformant les termes saisis par l'utilisateur en requête au format SQL¹⁶. En d'autres termes, l'ORM que nous utilisons, *SQL-Alchemy*

15. Voir commentaire du code et fichiers « donnees.py »

16. Le *Structured Query Language* est un langage de requête, modification et construction des bases de données relationnelles

permet de « traduire » des requêtes saisies en langage naturel au sein de la barre de recherche, en requêtes permettant de récupérer un ensemble de données dans une base de données. Ainsi, le moteur de recherche développé permet une recherche simple mais n'inclut pas pour le moment la possibilité de recherche plein texte. Dans cette optique, une possibilité de développement serait de passer d'une recherche par l'ORM à une recherche par indexation par le biais d'un moteur de type *ElasticSearch*¹⁷.

Dans un second temps, nous avons procédé à une indexation des toponymes et des organisations ou institutions suivant les recommandations de l'équipe scientifique et technique. D'un point de vue technique, nous avons recours à une feuille de transformation XSLT qui liste les éléments à indexer et automatise la création des liens vers les notices concernées. Nous avons développé ce même système d'indexation des dates afin de mettre en évidence les possibilités techniques de l'application web auprès de l'équipe scientifique. La mise en forme de cette chronologie a été réalisée exclusivement en CSS afin de palier l'absence de classe adaptée dans Bootstrap.

6.2.3 Une application pour les lier tous : le site web comme confluence des axes du projet

Au sein du projet, l'application web du *Dictionnaire numérique de la Ferme générale* tend à se placer à la confluence des 4 axes du projet. En effet, en tant que site web, le dictionnaire se propose d'inclure non seulement les notices, mais aussi l'ensemble des informations propres au développement du projet, incluant les publications, les renvois vers l'Atlas, la présentation du projet et de l'équipe, ainsi que la documentation et l'ouverture des données. Nous proposons d'expliquer la manière dont nous abordons ces différents enjeux en nous référant aux pages et onglets suivants :

- L'onglet « A propos » contient l'ensemble des informations relatives aux objectifs du projet, à son organisation par axes et à son équipe. La page « présentation du projet » reprend le texte d'introduction au *Dictionnaire de la Ferme générale* rédigé par Marie-Laure Legay initialement pour le carnet Hypotheses.org. Ce texte pourra faire l'objet d'une éditorialisation par la suite. La page « L'équipe du projet » présente les différents membres du projet répartis par domaine et axes de travail. Nous avons de même reproduis la liste des contributeurs aux notices scientifiques.
- L'onglet « Découvrir le projet » reprend les 4 axes du projet et contient l'index

17. La documentation officielle de ce moteur de recherche avancée se trouve à l'adresse suivante : <https://elasticsearch-py.readthedocs.io/en/v8.3.3/>

des notices intégrées, une page vierge devant intégrer dans le futur une partie ou l'entièreté de l'*Atlas de la Ferme générale* en cours de constitution, une page correspondant à l'Axe 3 reprenant l'ensemble des actualités du projet et des publications et une page contenant la documentation et les liens vers le code ouvert du projet.

- L'onglet « Rechercher » contient le moteur de recherche simple que nous avons décrit précédemment.
- Finalement, l'onglet « Index et chronologie » contient les différents index des entités nommées que nous souhaitons recenser.

Pour conclure cette présentation de l'application web, nous voudrions mettre en exergue un enjeu majeur auquel nous avons été confronté au cours du développement : la temporalité du projet. Effectivement, ce prototype s'inscrit dans un cadre un *workflow* global où les notices ne sont pas toutes rédigées et où l'Atlas n'est encore qu'à sa phase de développement. Dès lors, il nous faut penser « au-delà » du prototype pour ne pas fermer de portes à de futurs développements techniques et proposer une interopérabilité des informations, notamment avec les données géographiques de l'Atlas.

6.3 Au delà du prototype : perspectives de développement de l'application web

6.3.1 De l'échantillon au Dictionnaire : l'intégration en continu des notices

Dans un dernier temps, il nous faut mettre en évidence les éléments et fonctionnalités qu'il reste à implémenter au sein du prototype. Nous proposons d'analyser trois points clés de la mise en forme du dictionnaire qu'il faudra approfondir :

- Comme nous l'avons précédemment mentionné, ce prototype démonstratif n'inclut que les 29 notices de la lettre A. Un premier axe et enjeu de développement à venir est l'intégration en continu des notices. Ce passage de l'échelle de l'échantillon au dictionnaire implique de penser l'articulation et l'indexation des notices dans leur ensemble. Ainsi, il serait nécessaire d'amplifier les métadonnées décrivant les textes scientifiques afin de les catégoriser par lettres d'une part, puis par domaines d'analyse (histoire politique, judiciaire, financière,...) d'autre part. Dans cette optique, le prototype pourrait inclure une navigation plus approfondie

au sein du dictionnaire et un système de recherche avancé. L'enjeu sous-jacent étant d'accompagner et de suivre le développement scientifique du dictionnaire.

- Dans cette même optique de transition vers un dictionnaire complet, il serait indispensable d'implémenter le renvoi vers les notices citées ou liées au sein du texte. Pour ce faire, nous proposons à ce stade une liste des notices concernées qui seront par la suite implémentées dans le dictionnaire. Notons qu'un processus de normalisation des liens et urls devrait être développé.
- Une autre piste de développement serait de proposer une visualisation du dictionnaire sous forme d'un graphe afin d'une part de faire surgir les liens entre les notices au niveau global du dictionnaire, et d'autre part d'améliorer l'expérience utilisateur de navigation. Nous pouvons fournir comme élément de comparaison la visualisation en réseau proposée *l'Ontologie du christianisme médiéval en images* qui permet de mettre en exergue les liens sémantiques entre les différentes notices¹⁸.

6.3.2 La carte et le texte : anticiper les liens avec l'*Atlas de la Ferme générale*

Un autre enjeu majeur des développements du *Dictionnaire numérique de la Ferme générale* à venir est l'intégration et le croisement des données géographique produites par l'axe du projet en charge de la réalisation de l'*Atlas de la Ferme générale*. Si au moment de l'écriture de ces lignes la forme que prendra l'Atlas ainsi que le type de données ne sont pas encore connus¹⁹, nous pouvons tout de même anticiper l'importance de rendre les données interopérables. À cet égard, la mise en oeuvre de ce recueil de cartes est confié au CRESAT-UR 3436 (Centre de Recherche sur les Economies, les Sociétés, les Arts et Techniques) de l'Université de Haute-Alsace, sous la direction de Benjamin Furst. Nous pouvons donc nous baser sur les réalisations précédentes du CRESAT, ainsi que sur les indications transmises au cours des réunions inter-axes du projet. L'Atlas devrait prendre la forme d'une collection de cartes thématiques, aux échelles variées, permettant de spatialiser les enjeux de gouvernance ou de contestation étudiés au sein des notices²⁰. A titre de comparaison, nous pouvons citer l'*Atlas Historique d'Alsace* réalisé par le CRESAT qui propose une spatialisation d'un ensemble de thématiques géo-historiques relatives à

18. <https://omci.inha.fr/s/ocmi/page/omcigraph>

19. Le développement a proprement parler de l'Atlas doit commencer à partir de septembre 2022

20. Les objectifs de l'atlas sont ainsi formulés au sein de la présentation de l'ANR (<https://anr.fr/Projet-ANR-21-CE41-0019> : « Atlas. Il analyse la présence physique de la Ferme sur le territoire par province, mais aussi par bassin fluvial et par frontières. Il respecte l'esprit du projet qui souhaite, au-delà des données institutionnelles et judiciaires connues, établir des cartes qui représentent l'encadrement du privilège. »

l'Alsace, de l'antiquité à l'époque contemporaine²¹. La carte suivante traitant de l'enchevêtrément des souverainetés à l'aube de la Révolution française donne à voir un aperçu du type de cartes qui pourrait être produites dans le cadre de l'Atlas :

21. Voir <http://www.atlas.historique.alsace.uha.fr/fr/>

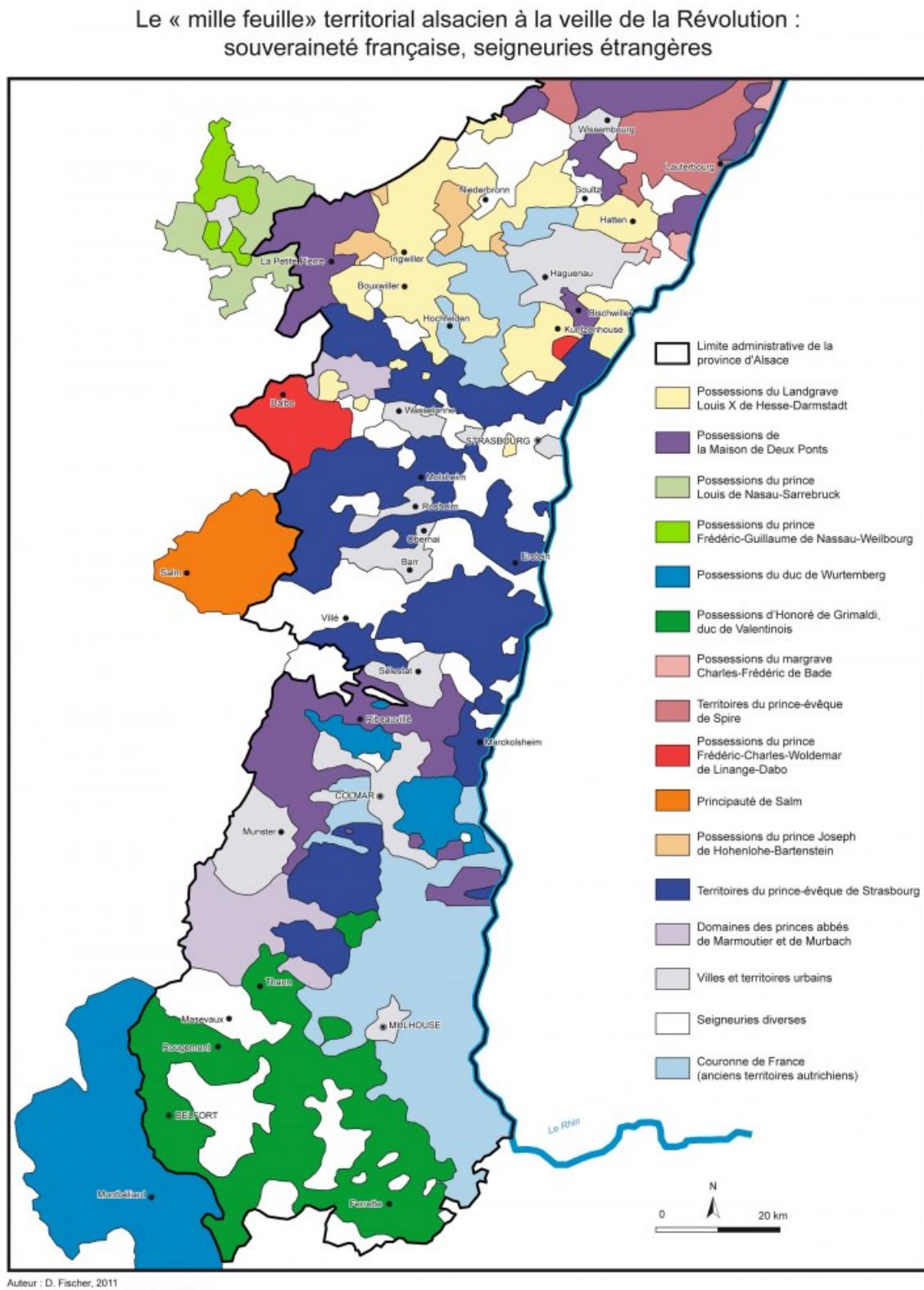


FIGURE 6.4 – Carte CRESAT - Daniel Fischer, « Le « mille feuille» territorial alsacien à la veille de la Révolution : souveraineté française, seigneuries étrangères », in Atlas historique d'Alsace, www.atlas.historique.alsace.uha.fr, Université de Haute Alsace, 2011

Dès lors, comment pouvons-nous croiser efficacement et de manière pérenne les don-

nées ou métadonnées majoritairement textuelles du Dictionnaire avec les données spatiales de l’Atlas ? Si ce travail n’a pu être qu’envisagé au cours du stage, nous pouvons tout de même fournir quelques réflexions et éléments de réponses :

- Un premier niveau d’intégration et croisement des données serait de produire un ensemble d’hyperliens renvoyant de part et d’autre de l’atlas ou du dictionnaire vers les notices ou les cartes associées. Cela impliquerait de développer, comme nous l’avons suggéré précédemment pour l’amélioration du moteur de recherche, une typologie des notices et des cartes afin d’amplifier les métadonnées et ajouter une couche sémantique aux diverses productions au sein du projet. Cette hypothèses impliquerait par exemple l’ajout d’un ensemble de mot-clés directement dans l’encodage en XML/TEI pour les notices du Dictionnaire. Cette solution est tout à fait envisageable puisque au moins deux balises TEI acceptent des listes de mot-clés ou des ontologies étrangères, à savoir les balises <xenoData²²> ou <keywords²³>. Cette solution est relativement facile à mettre en place mais implique de retravailler *a posteriori* les données produites.
- Un second niveau serait de conceptualiser et de mettre en place une interopérabilité maximale des données produites dès le début de la confection des cartes. Effectivement, les données géographiques peuvent être structurées autour de deux modèles principaux, soit en JSON (*JavaScript Object Notation*²⁴) proposant une arborescence sous formes de dictionnaires associant une clé et des valeurs, soit en KML (*Keyhole Markup Language*²⁵), à savoir un métalangage XML. Le format KML étant dérivé du XML, il serait possible de mettre en place un schéma XML afin de standardiser les données produites et ainsi automatiser par la suite le croisement des informations.

En tout état de cause, ces deux hypothèses impliquent une réflexion et mise en œuvre commune entre les axes du projet qui ont leur temporalité de recherche et développement propre. A cet égard, dans cette même optique de la temporalité longue du projet, un dernier point doit être aborder : l’hébergement de l’application web finale et l’ouverture des données au public.

22. <https://tei-c.org/release/doc/tei-p5-doc/en/html/ref-xenoData.html>

23. <https://tei-c.org/release/doc/tei-p5-doc/en/html/ref-keywords.html>

24. <https://json.org/json-fr.html>

25. Format *opensource* de structuration des données géographiques, principalement utilisé dans les SIG et maintenu par l’*Open Geospatial Consortium*. Voir <https://www.ogc.org/pressroom/pressreleases/857>

6.3.3 Passer à l'échelle : vers l'hébergement et l'ouverture de l'application

Finalement, un dernier point à aborder qui touche la phase finale du développement technique de l'axe 1 du dictionnaire est celui de l'hébergement futur du site web et de la mise à disposition des données. Comme nous l'avons précédemment mentionné, le projet ANR s'appuie sur les outils et infrastructures développés par HumaNum. Dans cette même optique, l'I.R* met à disposition des serveurs pour les projets relevant des humanités numériques au sein de l'Enseignement Supérieur et de la Recherche. Cette mise en place de l'hébergement par HumaNum présente une solution pérenne de mise à disposition des données. Néanmoins, plusieurs points techniques de gestion des données et métadonnées en lien avec la mise à disposition d'un serveur HumaNum doivent être soulignés. Nous nous réferrons dans ce cas aux conseils préalables et à la documentation fournie par l'IR sur ce sujet²⁶.

Tout d'abord, l'hébergement sur un serveur HumaNum implique de se conformer aux principes de l'interopérabilité et du signalement des données conforme au moissonnage mis en place par la plateforme ISIDORE. ISIDORE se présente sous la forme d'un moteur de recherche recensant les publications et données numériques dans le domaine des Sciences Humaines et Sociales. Ainsi, cette centralisation des données au sein de la plateforme s'appuie sur le protocole OAI-PMH (*Open Archives Initiative Protocol for Metadata Harvesting*), à savoir le protocole informatique développé dans le cadre de l'Open Archives Initiatives permettant d'échanger et centraliser les métadonnées au sein d'entre-pôts de données ou de moteurs de recherche institutionnels. Afin de rendre le site d'accueil du dictionnaire « moissonnable » et donc facilement trouvable par les chercheurs et futurs utilisateurs, le protocole requiert d'avoir accès aux métadonnées au format Dublin Core ou à d'autres schémas de données relevant du XML. La spécificité de cette mise en conformité pourrait donc impliquer la traduction de certaines métadonnées au format Dublin Core.

De plus, d'un point de vue technique, les serveurs d'hébergement reposent sur une architecture développée en langage PHP et sur des logiciels de base de données relationnelle en MySQL et en BaseX ou eXistDB pour les données XML. Dès lors, un dialogue en amont avec les ingénieurs d'HumaNum devra être mis en place afin de faciliter l'hébergement et l'interopérabilité des données.

En conclusion, le développement de ce prototype d'application web, laissant en-

26. La documentation se trouve à l'adresse suivante : <https://documentation.huma-num.fr/hebergement-web/>

trevoir la forme que pourrait prendre le *Dictionnaire numérique de la Ferme générale* a impliqué des choix techniques relevant de l'ouverture et de la pérennité des données. Effectivement, en choisissant un langage de programmation avec un cadre de travail très largement répandu (Python et Flask) au détriment d'un CMS de type Omeka, nous optons pour une inscription dans une très large perspective d'amélioration future. Que ce soit au niveau des fonctionnalités (moteur de recherche « plein texte », recherche à facette, lien avec des référentiels pérennes) ou de la mise à disposition des données (format XML/TEI), les choix effectués tentent de concilier d'une part les besoins concrets du projet et de l'équipe de recherche, et d'autre part les impératifs d'ouverture et pérennisation des données produites. De même, en proposant un code ouvert et documenté, nous mettons ainsi en pratique les principes de la « science ouverte » qui invitent à penser au delà du Dictionnaire et au cycle de vie de la donnée.

Chapitre 7

Vers le Dictionnaire et au-delà : ouvrir les données

Dans une dernière analyse, il nous semble nécessaire d'ouvrir notre perspective et de proposer une réflexion concernant le « cycle de vie » des données produites et la nécessité de s'inscrire dans le domaine de la science ouverte. Cette phase d'ouverture et de mise à disposition des données se trouve donc en aval de notre *workflow* que nous illustrons une dernière fois à la fin de cette présente introduction. Soulignons que l'inscription de notre travail et a fortiori du projet dans le mouvement de la science ouverte relève à la fois d'une obligation scientifique, puisque le quatrième axe du projet ANR est identifié comme assurant l'ouverture des données, et d'un ensemble de bonnes pratiques assurant l'interopérabilité et la pérennisation du code et des données produites. L'affiliation du projet à la « science ouverte » soulève donc la question de la nature même, en France, de ce mouvement d'ouverture des données au sein de l'Enseignement Supérieur et de la Recherche. À la fois obligation légale développée par deux plans nationaux depuis 2018 et pratique historique de partage des savoirs et techniques, la « science ouverte » se trouve indéniablement à la confluence des enjeux contemporains entourant la gestion des données scientifiques. Au-delà de cette analyse de la place et des limites de la science ouverte à l'échelle institutionnelle, il nous faut mettre en évidence la manière dont cet ensemble d'obligations et de recommandations est abordée au cours du projet. Ainsi, il nous faut nous demander finalement en quoi l'inscription du *Dictionnaire numérique de la Ferme générale* dans la « science ouverte » se place à la confluence des enjeux techniques, scientifiques et institutionnels entourant la vie de la donnée. Dès lors, nous proposons de terminer notre travail d'analyse en soulignant le fonctionnement théorique de la science ouverte et de ses outils, afin de mieux saisir leur mise en pratique au sein du projet et les possibles perspectives de développement à venir.

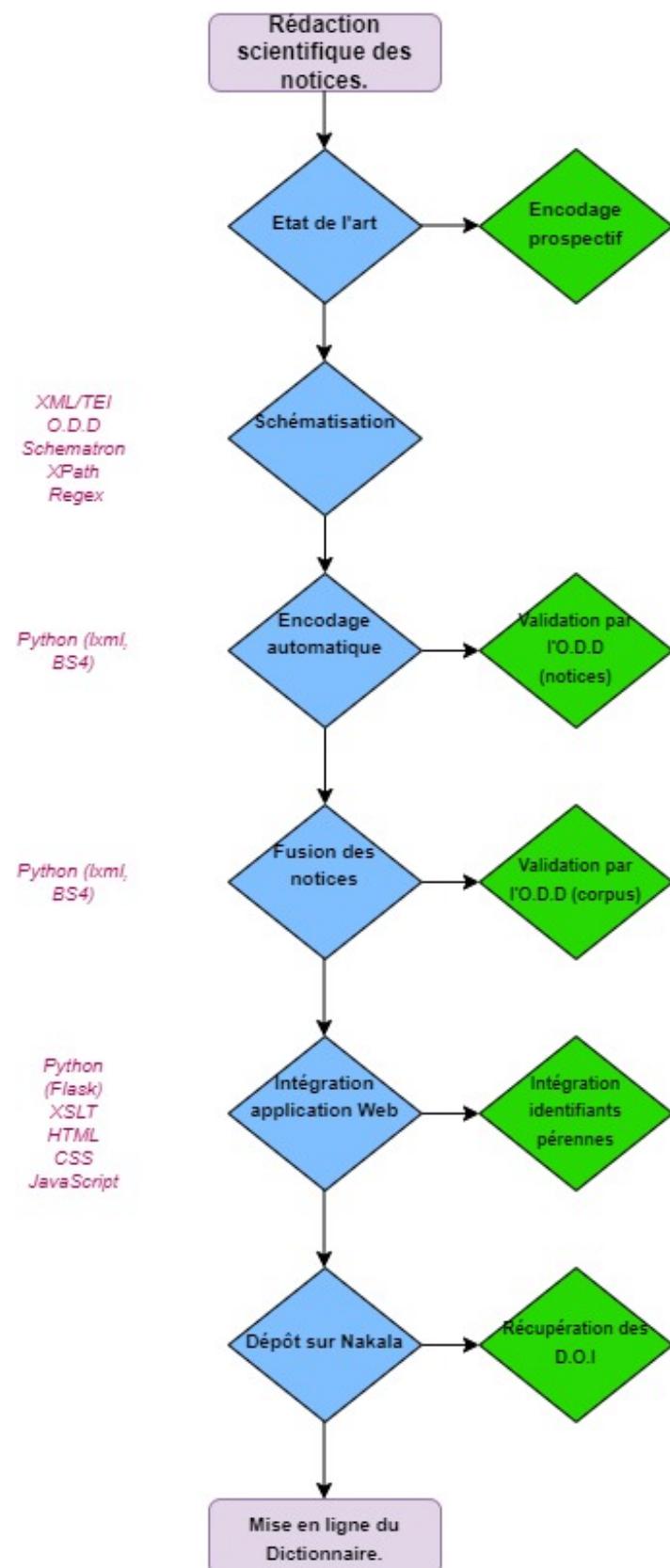


FIGURE 7.1 – *Workflow* : phase d’ouverture et mise à disposition des données.

7.1 Le Dictionnaire et la « science ouverte » : cadre théorique et mise en pratique

7.1.1 La « science ouverte » : un nouveau paradigme pour la recherche en sciences humaines et sociales ?

La « science ouverte » relève d'enjeux et d'ambitions scientifiques, techniques, économiques et, dans une moindre mesure, politiques. Il s'agit dès lors d'un mouvement plus que d'une école de pensée ou d'un courant de recherche, qui se caractérise par la diffusion sans entrave des publications et données de la recherche, le développement d'un écosystème de la donnée ouverte et une démocratisation de l'accès aux données, devant restaurer une forme de confiance entre le citoyens et les administrations¹. Ces grands enjeux sont définis dans le premier *Plan National pour la Science Ouverte* publié en Juillet 2018 par le Ministère de l'Enseignement Supérieur et de la Recherche, qui préconise dès lors trois axes de développement :

- Une généralisation de l'accès aux publications scientifiques, par le biais notamment du dépôt de publication HAL. L'objectif est de passer progressivement d'un paradigme quantitatif dans les publications scientifiques à un paradigme qualitatif. Afin de maintenir ces objectifs, la décision de développer le modèle « diamant » au sein des publications scientifiques a été prise, à savoir le financement par l'Etat des revues en accès ouvert². Dans cette même perspective, l'obligation de rendre accessibles et gratuits les articles publiés dans le cadre des projets ANR financés a été imposée en 2018 et renforcée en 2021 par la généralisation des Plans de Gestions de Données (PGD ou *Data Management Plan*), que nous évoquons plus bas.
- Une structuration des données produites dans le cadre de la recherche par la mise en conformité avec les principes FAIR (Facile à trouver, Accessible, Interopérable et Réutilisable). Pour ce faire, la création en 2022 du dépôt de données pluridisciplinaire Recherche Data Gouv, devant compléter plus que concurrencer les entrepôts disciplinaires, a été inauguré.
- Une insertion multi-scalaire dans une dynamique européenne et mondiale durable par l'inscription de laboratoires et institutions françaises dans les organismes de

1. Voir <https://www.ouvrirlascience.fr/plan-national-pour-la-science-ouverte/>

2. Dans ce modèle, l'Etat par le biais des consortiums et d'une enveloppe budgétaire spécifique prend en charge le financement des revues scientifiques en accès ouvert. Cela implique un renoncement à la rétribution des droits d'auteurs. L'objectif fixé pour l'horizon 2030 et d'atteindre un seuil de 75 % des revues en accès ouvert

coordination de la science ouverte, à l'instar de l'*European Open Science Cloud*.

Notons que ce premier plan, lancé en 2018, a été renouvelé par un *Deuxième Plan national pour la science ouverte* pour la période 2021-2024³. A cette occasion, un bilan fut dressé et de nouvelles recommandations furent formulées touchant la promotion des codes et logiciels libres, ainsi que la généralisation des pratiques de la science ouverte. Soulignons donc que depuis 2021, la focale semble s'élargir à l'ensemble des types de données pouvant être produits au cours d'un projet en humanités numériques tel que l'ANR FermeGé.

7.1.2 La « science ouverte » : une pratique historique de partage des savoirs et techniques

Afin d'apporter une vision plus englobante de la science ouverte, il nous faut nous détacher un tant soit peu de la perspective institutionnelle proposée par les plans nationaux pour la science ouverte. Notre objectif étant de souligner par la suite la manière dont l'inscription du projet dans la science ouverte est mise en oeuvre. À bien des égards, la science ouverte repose aussi pour partie sur une pratique historique de la mise en commun des savoirs et d'une véritable culture de l'accès libre qui s'est progressivement institutionnalisée. Comme le souligne Marin Dacos, ingénieur de recherche au CNRS, directeur du Centre pour l'édition électronique ouverte et, depuis 2017, conseiller scientifique du directeur général de la recherche et de l'innovation au Ministère de l'Enseignement Supérieur, de la recherche et de l'innovation, la science ouverte s'appuie sur une longue histoire du partage au sein de la communauté scientifique. Les promoteurs et adeptes de la science ouverte sont véritablement des « nains sur les épaules de géants »⁴. Le chercheur esquisse à grands traits l'historique des pratiques de la science ouverte, qui plonge ses racines dans le développement des revues scientifiques comme nouveau mode de diffusion des savoirs à partir de 1665 et l'émergence du *Journal des sçavans*. Cette revue des savoirs scientifiques se présente sous forme d'un bulletin visant à introduire et diffuser les informations concernant les inventions scientifiques et littéraires. Il s'agit donc d'un nouveau paradigme qui rompt avec la pratique de l'échange épistolaire traditionnel et participe donc au développement d'une pratique d'ouverture des savoirs qui s'accentue au cours des XIX^e et XX^e siècles⁵. Ce sont les dernières décennies du XX^e siècle et les premières du XXI^e qui voient une plus large diffusion et une structuration de la science ouverte au sein de la communauté scientifique. A cet égard, Marin Dacos distingue trois périodes décisives :

3. <https://www.enseignementsup-recherche.gouv.fr/fr/le-plan-national-pour-la-science-ouverte-2021-2024>

4. Dacos (Marin), « Des nains sur les épaules de géants: ouvrir la science en France », Revue Politique et Parlementaire (2019). URL : <https://hal.archives-ouvertes.fr/hal-02366604..> Consulté le 18 juillet 2022. p. 1 et 2

5. *Ibid.* p.4

- De 1991, date de la première expérience numérique d'archive ouverte pour les physiciens avec la plateforme ArXiv, jusque vers 2002-2003, cette décennie est marquée par un pragmatisme et foisonnement des pratiques. Par exemple, un système d'envoi par e-mail des articles scientifiques au format PDF est mis au point dans la communauté des biologistes américains. Cette première période voit les premiers jalons de la science ouverte se poser avec la fondation en 1999 de la première revue en accès ouverte, Revues.org qui devint par la suite OpenEdition.
- La seconde décennie de 2002 à 2010 se veut plus politique avec une prise de position de différents groupes de chercheurs par le biais des pétitions. Par exemple, la *Déclaration de Budapest* formulée en 2002 énonce les principes contemporains de l'accès ouvert et entend fédérer une communauté scientifique militante.
- Une troisième période depuis 2012 recoupe le principe d'institutionnalisation de la science ouverte que nous avons mis en évidence précédemment. En effet, la science ouverte est inscrite à l'agenda des politiques nationales et internationales de recherche. Ce mouvement s'est développé dès 2012 avec un ensemble de recommandations fournies par la Commission Européenne aux États membres à l'égard de la nécessaire ouverture des publications financées par des fonds européens ou nationaux⁶. C'est à partir de ces recommandations que ceux-ci mirent en place les plans nationaux pour la science ouverte, à l'instar des Pays-Bas dès 2016 ou de la France en 2018.

En conséquence, la science ouverte s'appuie sur une pratique de partage de connaissances et des techniques qui est partie prenante du travail des chercheurs au travers des publications ouvertes. Ces pratiques communautaires se structurent progressivement et s'institutionnalisent dans le cadre des projets de recherche nationaux, à l'instar de l'ANR FermeGé, par les recommandations et plans nationaux de la science ouverte depuis 2018. Ce cadre notionnel à présent posé, nous pouvons mettre en évidence la manière dont les principes de la science ouverte sont appliqués au sein du projet.

6. *Ibid.* p.8. La commission dressait alors un constat alarmant sur la perte économique que représente le modèle des revues scientifiques payantes mais financées par des fonds publics. Avec un investissement de 80 milliards d'euros dépensés par an, les retombées économiques n'étaient, d'après la Commission, que très peu perceptibles. Remarquons donc que l'enjeu économique n'est pas absent des préoccupations de la science ouverte, que l'on retrouve aussi dans la notion de mutualisation des outils.

7.1.3 La science ouverte et ses outils : le cas du *Dictionnaire Numérique de la Ferme générale*

Dans le cadre de l'axe 1 du projet ANR FermeGé consacré à l'élaboration scientifique et technique du dictionnaire, les principes de la science ouverte sont mis en pratique autour de trois points relatifs au cycle de vie de la donnée : l'encodage dans des formats libres, non propriétaires et interopérables ; la gestion des données au niveau institutionnel par un Plan des Gestion des données et finalement l'archivage des données dans un dépôt en accès libre. Le schéma suivant entend illustrer l'agencement et la complémentarité de ces différents outils :

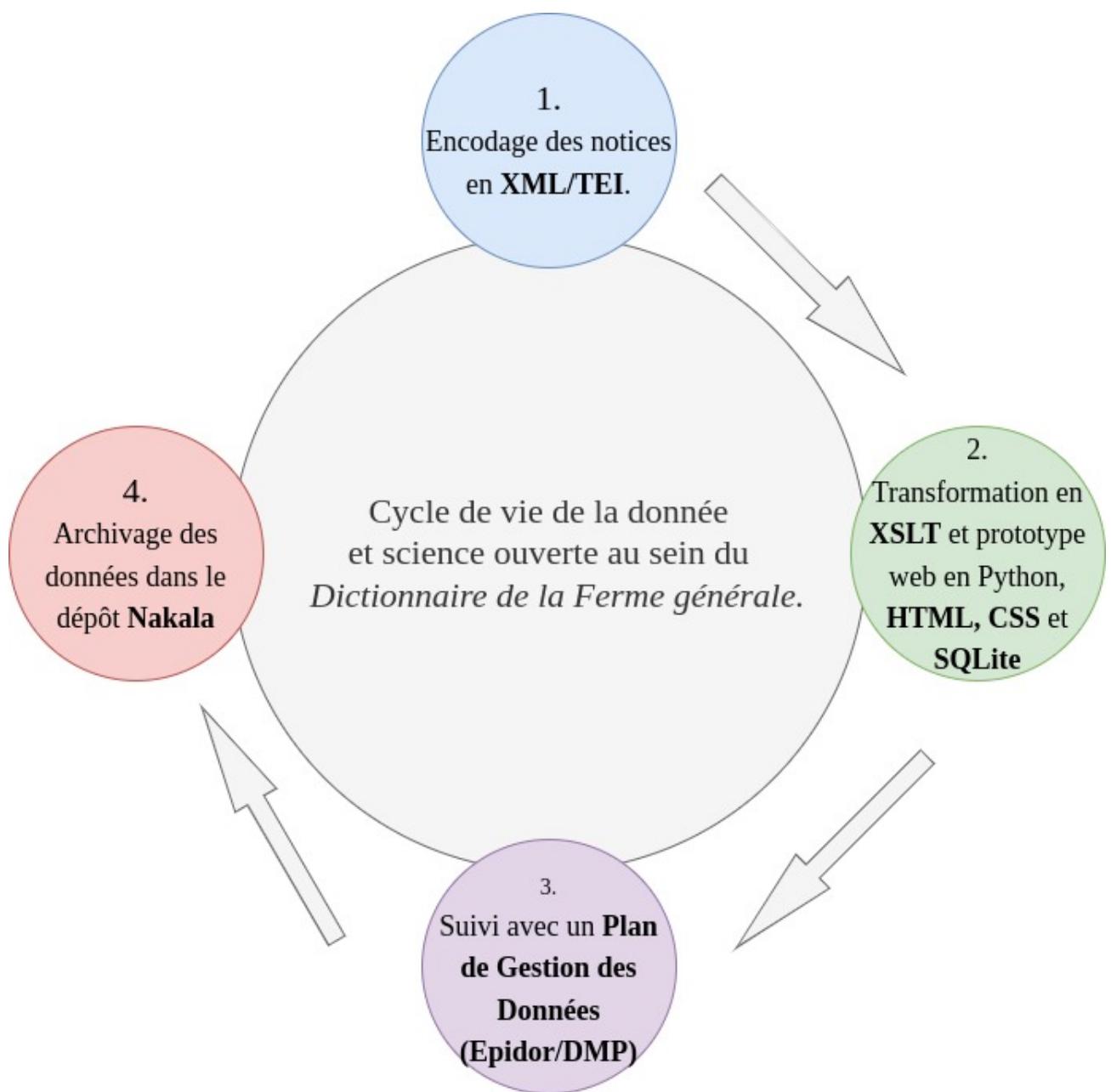


FIGURE 7.2 – Schéma : cycle de vie de la donnée et science ouverte dans le projet FermeGé.

Comme nous l'avons précédemment évoqué, les formats et outils de développement pour automatiser l'encodage des notices s'inscrivent pleinement dans le mouvement de la science ouverte, puisque nous utilisons des formats interopérables tels que l'XML/TEI pour l'encodage ou l'XSLT (eXtensible Stylesheet Language Transformations) pour la transformation des données. Afin d'assurer une certaine pérennité du projet, cet encodage a été documenté et explicité dans la première partie « documentation » de l'O.D.D. De même, l'ensemble de notre code a été commenté et rendu accessible à la communauté d'utilisateurs en créant un dépôt Gitlab accueillant les fichiers en question⁷. Ainsi, par cette pratique courante de mise à disposition du code commenté, nous nous inscrivons pleinement dans le nouvel axe d'ouverture des données au sein du Deuxième Plan national pour la science ouverte formulé en 2021, à savoir l'ouverture et l'archivage sur le long terme des codes et logiciels développés par des projets financés⁸.

De plus, l'ancrage dans la science ouverte s'affirme au sein du projet par la mise en place d'un Plan de Gestion des Données (PGD). Rédigé et mis à jour par Victoria Le Fournier et Marie-Laure Legay, ce document dématérialisé entend mettre en évidence la manière dont les données sont rassemblées, produites, gérées et archivées tout au long du projet. La plateforme utilisée est *OpidorDPM*⁹ puisqu'il s'agit du système de rédaction et gestion de PGD consacré à l'Enseignement Supérieur et la Recherche. Le plan est alimenté plusieurs fois au cours du projet à intervalles de quelques mois afin de pouvoir tracer la mise en place et l'évolution de la gestion et production des données. De manière très concrète le PGD est formalisé à partir de différents axes thématiques présentés sous forme de questions ouvertes traitant de l'acquisition et de la gestion des données, de leur traitement, des enjeux éthiques et juridiques s'il s'agit de données sensibles et leur diffusion puis archivage. L'objectif global est d'une part de démontrer l'inscription des données produites dans les principes FAIR (*Findable, Accessible, Interoperable, Reusable*) et d'autre part de « sensibiliser » les chercheurs à l'importance d'une gestion raisonnée et pérenne des données et métadonnées. Dans le cadre du projet, le PGD se structure autour des différentes tâches inhérentes aux axes du projet (gestion des données du dictionnaire, de l'atlas, des productions scientifiques). En somme, la rédaction d'un PGD s'inscrit dans les impératifs de transparence liés au financement des projets de recherche publics énoncé au sein des plans nationaux pour la science ouverte.

Finalement, en aval du cycle de vie de la donnée se trouve la question de l'archivage et de la publication des données. Dans l' « écosystème HumaNum », ces enjeux sont gérés par l'accès au dépôt de données numériques et de publications intitulé Nakala. Ce dépôt

7. Voir annexes - Dépôt Gitlab - 7.3.3. Lien vers le Gitlab du projet : <https://gitlab.huma-num.fr/vdecreaene/DicoNumFermeGe>

8. <https://www.enseignementsup-recherche.gouv.fr/fr/le-plan-national-pour-la-science-ouverte-2021>

9. <https://dmp.opidor.fr/>.

s’insère pleinement dans les principes et pratiques de la science ouverte puisqu’il permet à la communauté des chercheurs de partager, publier et valoriser un vaste ensemble de type de données issues des projets de recherche¹⁰. Les données doivent donc se conformer aux principes FAIR précédemment énoncés. Au-delà du stockage et de la publication, l’entrepôt Nakala prend en charge la génération de DOI afin d’assurer un système de citation pérenne pour chaque collection de données déposées. Ainsi, il sera possible à terme de déposer l’ensemble des notices afin de leur attribuer un identifiant que nous pourrons récupérer et ajouter au sein de la visualisation des notices proposées dans l’application web. Soulignons le fait qu’au sein de l’entrepôt Nakala, les métadonnées sont décrites en Dublin Core et renseignées au moment du dépôt manuel. Un nouvel enjeu est donc l’association et la traduction des métadonnées structurées en XML/TEI vers le Dublin Core étendu¹¹. En conséquence, le dépôt Nakala assure la phase finale du cycle de la donnée¹². Nous pouvons donc conclure que la science ouverte s’insère dans un ensemble de pratiques de développement et d’outils *opensource* permettant d’aider les équipes de chercheurs et d’ingénieurs dans l’ouverture et la gestion des données. Cependant, il reste un aspect relativement complexe en lien avec les enjeux de la science ouverte : les aspects juridiques ou éthiques, entourant la réutilisation des illustrations et numérisation des sources.

7.2 La science ouverte et les enjeux éthiques ou juridiques

7.2.1 Une diversité de documents et illustrations : quels enjeux juridiques ?

Si la majorité des données produites au sein du projet FermeGé sont textuelles, encodées en XML/TEI ou du code informatique, certaines données insérées dans les notices sont plus complexes à traiter et implique un recul critique afin d’en saisir les enjeux juridiques. Effectivement, en l’état actuel du *Dictionnaire de la Ferme générale*, divers documents iconographiques ont été insérés par les auteurs dont nous proposons ci-dessous une brève typologie :

- Schéma, graphiques et croquis produits directement par les auteurs des notices

10. <https://documentation.huma-num.fr/nakala/>

11. Le groupe de recherche « TeiNK » s’est constitué autour de cet enjeu technique et sémantique de *mapping* des métadonnées au sein de l’application

12. Indiquons de même que Nakala propose un module « Nakala Press » permettant une publication web relativement sommaire sur le modèle d’un CMS tel qu’Omeka. La problématique qui semble se dessiner concernant Nakala est sa capacité de fédérer ou non une communauté de chercheurs en SHS à l’instar d’Omeka ou Omeka S

à partir de données tirées de sources historiques. Ces graphiques sont souvent insérés directement au fil du texte des notices afin de proposer une visualisation d'un phénomène économique.

- Reproduction d'un document d'archive par le biais d'un cliché pris directement par l'auteur de la notice.
- Reproduction et insertion d'un document iconographique tiré de Gallica.fr. Il s'agit ici généralement de numérisation d'extraits d'un manuscrit conservé à la Bibliothèque nationale de France.
- La reproduction d'un document issu d'un ouvrage scientifique publié. Nous trouvons par exemple des illustrations et cartes tirées d'édition de sources primaires.

Dès lors, nous percevons tout l'enjeu et la tension au sein même des pratiques de la science ouverte entre le cadre juridique autorisant ou non la reproduction des documents et la nécessité de gérer les mentions légales au sein de l'application web. Il nous semble donc intéressant de poser brièvement le cadre juridique concernant ces divers documents afin de mettre en évidence la manière dont nous pouvons les intégrer au sein du *Dictionnaire de la Ferme générale*.

7.2.2 Réutiliser les documents dans le cadre de l'enseignement supérieur et de la recherche

Observons le panorama global du cadre juridique dans lequel s'inscrivent les différents documents que nous avons à notre disposition. Nous avons réalisé à cet égard un bref *memorandum* qui devra par la suite être vérifié et complété si nécessaire par un expert des questions juridiques de réutilisation des données. Nous pouvons à cet égard esquisser le cadre normatif et les droits de réutilisation dont nous disposons au sein du projet.

Tout d'abord, la reproduction des documents d'archives et se trouvant sur Gallica est encadrée par la loi du 28 décembre 2015 relative aux modalités de la réutilisation des informations du secteur public, dite loi Valter¹³ qui transpose dans le droit français la directive européenne du 26 juin 2013 sur le même sujet. Comme le souligne le juriste Bruno Ricard dans son article intitulé « Le nouveau régime juridique de la réutilisation des informations publiques » pour la revue en ligne *Droit(s) des archives*, cette loi vient dans une certaine mesure compléter le code du patrimoine de 2004 sur la question des

13. <https://www.legifrance.gouv.fr/loda/id/JORFTEXT000031701525/>

données issues des documents d'archive¹⁴. Ainsi, dans le cadre de cette loi, la libre réutilisation des informations publiques est affirmée à condition que « ces dernières ne soient pas altérées, que leur sens ne soit pas dénaturé et que leurs sources et la date de leur dernière mise à jour soient mentionnées ». Par conséquent, la diffusion de reproduction de documents d'archives ne semble pas poser de problème juridique. Les documents accessibles sur Gallica sont régis par cette même loi et autorise une reproduction non commerciale¹⁵. Les métadonnées de ces documents sont soumises à la licence EtaLab et autorise donc un usage libre et gratuit sous réserve de mention de la source originale.

Le cas des graphiques, schémas et autres croquis produits par les auteurs des notices relèvent du *Code de la propriété intellectuel* du 3 juillet 1992¹⁶. Il permet à l'auteur d'une œuvre de décider de la manière dont son œuvre peut être diffusée et utilisée. Ainsi, la demande formelle d'autorisation de reproduction est nécessaire.

Le cas des reproductions de documents iconographiques tirés des ouvrages scientifiques est plus complexe. Effectivement, les documents de recherche internes au projet relèvent de l' « exception pédagogique et de recherche », la diffusion dans le cadre de l'application web implique un changement de paradigme juridique. Soulignons tout d'abord que l'exception pédagogique et de recherche relève de la loi de 2006 sur ce sujet qui autorise la reproduction et diffusion des œuvres et productions intellectuelles de manière illimitée dans le cadre de recherches communes ou de l'enseignement¹⁷. Cette exception repose sur le principe de compensation versée par l'État annuellement aux sociétés de gestion collective pour être ensuite reversée aux ayants droit. Néanmoins, dans le cadre de l'application web lorsque celle-ci sera déployée, ces mêmes documents tombent dans le cadre du Code de la propriété intellectuelle.

7.2.3 Outils et respect des enjeux éthiques et juridiques au sein du projet

Concluons cette mise en évidence des enjeux juridiques indissociables des pratiques de la science ouverte en mettant en exergue les propositions et solutions, certes partielles, que nous avons développées à cet égard. Le tableau suivant entend synthétiser les différents cas auxquels nous sommes confrontés.

14. Bruno Ricard, « Le nouveau régime juridique de la réutilisation des informations publiques », *Billet, Droit(s) des archives*, 2022. <https://siafdroit.hypotheses.org/659..> Consulté le 27 juillet 2022

15. « Conditions d'utilisation de Gallica » : <https://gallica.bnf.fr/edit/und/conditions-dutilisation-des-contenus-de-gallica>

16. <https://www.legifrance.gouv.fr/codes/id/LEGITEXT000006069414/>

17. Françoise Acquier, « Quelles images puis-je publier dans ma thèse ? Réponse en pratique dans un atelier pédagogique », *Billet, Ethique et droit*, 2019. URL : <https://ethiquedroit.hypotheses.org/2947..> Consulté le 27 juillet 2022.

Type de document.	Cadre législatif.	Gestion des mentions légales.
1. Reproduction de documents d'archives	Loi Valter du 28 décembre 2015 relative à la gratuité et aux modalités de réutilisation des informations du secteur public	Demander une autorisation de reproduction du cliché à l'auteur.
2. Reproduction d'un document tiré de Gallica.fr	Loi Valter du 28 décembre 2015. Licence EtaLab	Mentionner systématiquement la source du document.
3. Document produit en interne au projet (graphiques, schémas et cartes).	Code de la propriété intellectuelle (3 juillet 1992)	Demander une autorisation de reproduction du cliché à l'auteur.
4. Cliché ou reproduction d'un document issu d'un ouvrage publié sous licence libre.	Code de la propriété intellectuelle (3 juillet 1992)	Si l'ouvrage est sous licence libre (CC BY), reproduction des mentions légales de la licence.
5. Cliché ou reproduction d'un document issu d'un ouvrage publié sans licence libre.	Code de la propriété intellectuelle (3 juillet 1992)	Si l'ouvrage est édité, envoyer une demande explicite d'autorisation au titulaire du droit d'auteur.

FIGURE 7.3 – Tableau : synthèse des enjeux juridiques par type de documents.

Dans l'état actuel du prototype de l'application web, nous avons procédé d'une part à un recensement des documents iconographiques présents dans les notices, puis nous avons ajouté une page « mentions légales » présentant la source de chacun des documents en question¹⁸. En somme, la pratique de la science ouverte s'articule aussi autour d'enjeux juridiques qui peuvent évoluer au cours du projet. Dans cette même perspective, nous pouvonsachever notre analyse en soulignant diverses possibilités de développement afin d'approfondir et anticiper une plus grande insertion dans la science ouverte.

18. Voir capture d'écran en annexes - Figure 22

7.3 Ouvrir les données du Dictionnaire : perspectives de développement

7.3.1 Développer une API : enjeux, processus et limites

Dans un dernier temps, il nous faut mettre en évidence diverses possibilités d'amélioration et d'approfondissement de l'ancrage du dictionnaire dans les principes et pratiques de la science ouverte. Dans cette perspective, nous pouvons soumettre l'idée que l'implémentation d'une API au sein de l'application web serait un garant de l'ouverture des données. Afin de mieux en saisir les enjeux techniques et les limites, il nous faut tout d'abord comprendre ce que sont les *Application Programming Interface* (API). Une API est un système informatique prévu pour la communication de données ou de fonctions à des services tiers. Cette interface permet donc le transfert vers un client externe des données dans un format brut sans passer par l'interface web¹⁹. Dans l'environnement web, les API prennent donc des formes diverses incluant par exemple des transferts d'images en très haute qualité comme le IIIF (*International Image Interoperability Framework*²⁰), des protocoles de transferts de métadonnées tel que l'OAI-PMH évoqué plus haut, ou des fonctions web plus complexes comme les applications de traduction instantanées de *Google*. Dans le cadre des humanités numériques et de la science ouverte, l'enjeu est donc la construction de cette interface de communication des données dans des formats ouverts et non propriétaires. Afin de répondre à ces enjeux de qualités des données transmises, il est nécessaire pour une API de se conformer aux spécifications REST (*Representational State Transfer*) énoncés par Roy Fielding dans sa thèse pionnière portant sur la conceptualisation de l'architecture des API²¹. D'après lui, un ensemble de principes techniques régissent la bonne construction d'une API, dont trois nous intéressent tout particulièrement dans le cadre d'un projet en humanités : l'inscription du fonctionnement de l'API dans les principes du protocole web HTTP²², une interface uniforme impliquant l'accès à des données dans un format structuré et ouvert²³ et le « système par échelle (*layered system*) qui permet le traitement de requêtes en masse. En somme, le développement d'une API serait un moyen d'ouvrir l'accès aux données du dictionnaire et à ses métadonnées à une plus grande échelle. Effectivement, par la possibilité de requêter en masse des métadonnées dans un format libre et intéropérable comme le JSON, cela permettrait d'approfondir l'ancrage du projet dans la science ouverte.

19. Roy T. Fielding, *Architectural styles and the design of network-based software architectures*, University of California, 2000.

20. <https://iiif.io/api/index.html>

21. *Ibid.*, voir chapitre 5 *Representational State Transfer (REST)*

22. Cette spécification évidente permet une meilleur insertion dans le web de données et par extension dans la science ouverte

23. La majeur partie des API modernes proposent l'accès à des données en masse au format JSON ou CSV

D'un point de vue technique au sein de l'application web, le développement d'une API consacrée est facilitée par le *framework* Flask que nous utilisons. Avant d'entrer plus en amont dans des hypothèses de développement technique, nous pouvons apporter des éléments de comparaison en terme de mise en place d'API. En premier lieu, la Bibliothèque nationale de France se trouve à la pointe du partage des métadonnées par le biais des API de Gallica. Effectivement, la BnF propose cinq API donnant accès aux données et métadonnées des catalogues des livres numériques²⁴, aux images en hautes définition au format IIIF²⁵, à l'entrepôt OAI-PMH²⁶, aux données de recherche sur Gallica²⁷ et aux informations nécessaires à l'exploitation des ressources numériques²⁸. Le cas des API de la BnF est tout particulièrement intéressant en raison de la diversité des données accessibles. Néanmoins force est de constater que leur utilisation requiert une certaine maîtrise technique des protocoles de requêtes et une connaissance des formats de données renvoyées. A titre comparatif, l'API développée au sein du projet du *Dictionnaire topographique de la France* soutenu par l'École Nationale des Chartes, le Comité des Travaux Historiques et Scientifiques (CTHS) et les Archives Nationales (AN) est tout à fait remarquable²⁹. En effet, elle fournit des réponses aux requêtes concernant des toponymes au format JSON associant les données et métadonnées à des référentiels pérennes. Par exemple, les formes anciennes et contemporaines des toponymes sont croisées au sein de la réponse fournie par l'API, ce qui permet de récupérer le code INSEE de la commune et Geoname, l'identifiant Wikidata et ark de la BnF, ou les numéros VIAF et SIAF lorsqu'ils sont disponibles. En conséquence, nous comprenons que l'ouverture de ces données dans le cadre de la science ouverte ne peut se faire que par l'inscription du projet dans le web de données dont l'API est une facette.

Il est envisageable d'implémenter une API renvoyant des données au format JSON au sein de l'application web dans son état actuel. En effet, le format JSON permet de traiter de plus grandes quantités de données avec une plan grande performance. A cet égard, le tableau comparatif ci-dessous met en évidence les avantages de l'échange des données brutes dans un environnement web³⁰ :

24. <https://api.bnf.fr/fr/api-opds-du-catalogue-de-livres-numeriques-de-gallica>

25. <https://api.bnf.fr/fr/api-iiif-de-recuperation-des-images-de-gallica>

26. <https://api.bnf.fr/fr/node/170>

27. <https://api.bnf.fr/fr/api-gallica-de-recherche>

28. <https://api.bnf.fr/fr/api-document-de-gallica>

29. Voir <https://dicotopo.cths.fr/documentation>

30. Tableau adapté de l'article suivant : Saurabh Zunke et Veronica D'Souza, « JSON vs XML : A Comparative Performance Analysis of Data Exchange Formats », *International Journal of Computer Science and Network*, vol. 3, n° 4 (2014), p. 5.

Élément de comparaison	HTML	JSON
Poids	7364 o	5257 o (mais avec plus d'informations propres à la requête)
Parsage	Capable d'échouer si une page web est mal construire dans son header	Très peu probable d'échouer
Disponibilités dans les langages	Moyenne (librairies externes souvent)	Élevée

FIGURE 7.4 – Tableau comparatif des formats de données au sein d'une API.

Ainsi, afin de mettre en place cette API, il serait nécessaire de pouvoir traduire les données et métadonnées XML/TEI au format JSON. Ceci est réalisable en Python et dans le *framework* Flask par le biais d'un module intitulé « jsonify³¹ ». Si la mise en place d'une API est un aspect clé de l'ouverture des données et de leur inscription dans la science ouverte, cette solution n'est pas sans présenter certaines limites qu'il nous faut mentionner.

En effet, le développement d'une API ne peut pas être envisagé comme seul moyen d'assurer l'ouverture des données. C'est pourquoi il nous semble nécessaire d'insérer l'API dans une pratique plus globale de la science ouverte. Notons que l'utilisation d'une API n'est pas nécessairement aisée et peut présenter des difficultés de maintenance et de mise à jour. Il est de même relativement complexe d'anticiper l'utilisation des données de l'API que pourront en faire les chercheurs et utilisateurs. Sur ce point, certains utilisateurs soulignent les difficultés d'accès aux données par ce biais et les promesses parfois non tenues d'ouverture des données. La chercheuse et historienne de l'art Clarisse Bardiot dans son billet de recherche intitulé « Happy APIs : Débridons les APIS pour développer les humanités numériques » sur son carnet Hypothèses.org³² met en exergue un ensemble de limites liées à l'utilisation des API :

31. <https://flask.palletsprojects.com/en/2.2.x/api/?highlight=jsonify#flask.json.jsonify>

32. Clarisse Bardiot, « Happy APIs : Débridons les APIS pour développer les humanités numériques », Billet, DORRA-DH, 2018. URL : <https://dorradh.hypotheses.org/66..> Consulté le 19 juillet 2022.

- L'hétérogénéité du fonctionnement des API et de leur clés d'accès impliquant une appropriation parfois longue de leur fonctionnement.
- Le renvoi des données au format JSON entraînant une plus grande difficulté de réutilisation qu'un fichier au format CSV.
- Le manque de maintien et de réponses des institutions en charge des sites hébergeant des API.

Ces différentes critiques doivent indéniablement être prises en compte dans la phase de développement de l'API. Afin de palier ces potentiels problèmes, il est de même envisageable d'étoffer la page de l'application web contenant la documentation du projet³³ et d'y insérer une possibilité d'export au format CSV. En outre, l'inscription du projet dans la science ouverte peut aussi s'effectuer par la participation active du projet à ce mouvement et sa médiatisation.

7.3.2 Participer à la science ouverte : le cas des *data papers* et la « médiatisation » du projet

De prime abord, soulignons le fait que le mouvement de la science ouverte implique non seulement des changements de pratiques dans le développement informatique des outils, mais aussi dans la diffusion des connaissances et des résultats de la recherche. L'émergence récente des *data papers* en est sûrement le signe le plus visible. Les *data papers* sont des articles de revues scientifiques et leurs métadonnées décrivant un ensemble particulier de données produites dans le cadre d'un travail de recherche et publiés dans une revue académique traditionnelle basée sur le principe du *peer-review*³⁴. Ce nouveau type de publication symbolise un changement de paradigme qu'entraîne la diffusion de la science ouverte puisqu'ils s'intéressent aux données produites et à leur condition de production, plutôt qu'à leur analyse et critique scientifique. Ces *data papers* peuvent contenir des liens renvoyant vers des dépôts de données afin de participer à leur diffusion et citation. En somme, il s'agit autant d'un moyen de partager les métadonnées associées à un ensemble de données produites que de donner un forme « d'assise académique » aux personnes en charge de la publication et de leur mise à disposition³⁵. A cet égard, les

33. voir Annexes - Figure 20) ou onglet Découvrir le projet > Axe 4 : Un ancrage dans les principes de la science ouverte

34. Joachim Schöpfel, Dominic J. Farace, Hélène Prost, Antonella Zane, « Data papers as a new form of knowledge organization in the field of research data », *12ème Colloque international d'ISKO-France : Données et mégadonnées ouvertes en SHS : de nouveaux enjeux pour l'état et l'organisation des connaissances ?*, ISKO France, Oct 2019, Montpellier, France. halshs-02284548. p.4

35. *Ibid.*, p.5

chercheurs Joachim Schöpfel, Dominic J. Farace, Hélène Prost et Antonella Zane se sont intéressés à la publication des *data papers* dans les domaines de la biologie, de l'ingénierie, de la physique, de la chimie et de la biologie. Voici les conclusions tirées sur un corpus d'une centaine de revues :

- Les *data papers* permettent une généralisation de la structuration des métadonnées associées aux dépôts de données, assurant une meilleure automatisation de leur réutilisation³⁶.
- Ils permettent une structuration des critères d'évaluation des qualités d'un *dataset*³⁷.
- Ils impliquent néanmoins une charge de travail supplémentaire qui n'est encore que partiellement automatisable³⁸.

Le schéma suivant que nous tirons de ces recherches met en évidence le rapprochement de différents types de productions académiques propre à la science ouverte :

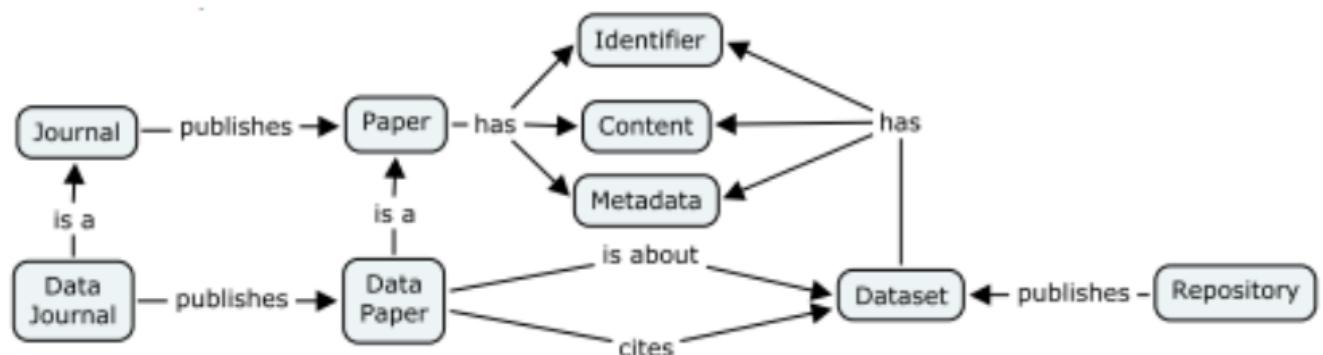


FIGURE 7.5 – Schéma illustrant l'insertion des *datapapers* dans l'écosystème des publications académiques.

En somme, la spécificité des *data papers* est qu'ils se situent à l'intersection entre publications scientifiques et dépôts brut de données. Cela nous invite donc à questionner les conséquences des données ouvertes au sein d'un projet en humanités numériques qui sépare traditionnellement les livrables techniques, les données produites et les analyses et conclusions scientifiques. Dès lors, nous pouvons nous questionner sur la pertinence et la possibilité de mise en oeuvre d'un *data paper* traitant du *Dictionnaire numérique de la*

36. *Ibid.*, p. 5 à 8

37. *Ibid.*, p. 11-12

38. *Ibid.*, p.14

Ferme générale.

Il nous semble qu'en l'état actuel du corpus à notre disposition, la rédaction d'un *data paper* apparaît certes comme une possibilité intéressante de participer activement à la science ouverte, mais présente plusieurs limites qu'il nous faut prendre en compte. Effectivement, comme nous l'avons précédemment suggéré, nous disposons certes de métadonnées structurées pour l'ensemble des notices, mais leur qualité peut encore être améliorée. Dans cette optique, la possibilité d'ajouter des mot-clés proposant une typologie des sujets traités ou l'incorporation d'une autre ontologie pourrait permettre de décrire plus en détail les jeux des données à notre disposition. De plus, les *data papers* sont-ils adaptés à la publication de données textuelles encodés dans un métalangage sémantique tel que la TEI? Il semble en effet que l'immense majorité des jeux de données analysés dans les *data papers* soient des données numériques, structurées dans des formats classiques et interopérables de type CSV. Finalement, la question de la quantité de données à notre disposition doit aussi se poser. Ainsi, il nous semble que la rédaction d'un *data paper* n'est pas le seul moyen de mettre en exergue les données produites dans le cadre du projet. En tout état de cause, la question de la diffusion des données produites dans le cadre du projet doit se poser sous l'angle de leur pertinence et de leur interopérabilité. Dans cette même perspective, nous voudrions dans une dernière analyse amener une réflexion sur le dictionnaire, et tout spécialement le travail que nous avons réalisé, en terme de modèle et conceptualisation des données.

7.3.3 Réflexions sur le Dictionnaire, ses données et leur utilisation

A l'aune de ce présent travail et des analyses que nous avons pu tirer du mouvement de la science ouverte, posons la question suivante : le *Dictionnaire de la Ferme générale* n'est-il pas *in fine* une base de données ? Derrière cette question incongrue se cache en réalité l'articulation entre sémantisme des données encodées en XML/TEI et leur utilisation future. Effectivement, nous ne reviendrons pas ici sur le modèle conceptuel appliqué aux notices qui justifie pleinement le qualificatif de dictionnaire associé à notre corpus. Rappelons simplement que notre corpus se structure indéniablement autour d'une dichotomie entre le lemme et son sens que la TEI vient caractériser. D'un point de vue de l'utilisation des données du corpus, celles-ci sont utilisées dans le cadre de l'application web comme une base de données. Le tableau synthétique suivant entend mettre en évidence les caractéristiques des données relevant d'une part du concept même de dictionnaire et d'autre part d'une base de données.

	Dictionnaire	Base de données document
Gestion des données textuelles	Ajout de sémantisme propre aux dictionnaires avec l'encodage automatisé en XML/TEI.	Données assemblées au sein d'un corpus requêtable comme une base de données documents (XPath et XQuery)
Gestion des métadonnées	Métadonnées renseignées dans l'en-tête TEI (<teiHeader>).	Ajout des métadonnées dans une base de données relationnelle en lien avec l'application web.
Proximité avec d'autre projet de recherche	Peu de similarités avec les dictionnaires linguistiques encodées en TEI.	Proximité avec projets ayant des corpus désignés comme "base de données" (Katabase, <i>Thesaurus Exemplorum Medii Aevi</i> , ect)
Inscription dans les principes de la science ouverte.	Mise à disposition des données sous forme de dépôt de fichier TEI	Développement d'une API.

FIGURE 7.6 – Tableau comparatif des éléments relevant d'un dictionnaire et d'une base de données.

En somme, cette analyse comparative renvoie à la question de la conception du *Dictionnaire de la Ferme générale* comme entité intellectuelle. S'agit-il d'une collection de notices reliées sémantiquement entre elles par des renvois ou une oeuvre organique à part entière ? Comme nous l'avons développé au cours des chapitres 4 et 5 de ce présent mémoire, nous avons décidé de proposer un moyen de concilier ces deux approches, pouvant répondre à des enjeux techniques spécifiques à diverses points du *workflow* du projet.

Pour conclure, l'inscription du projet FermeGé dans les principes et méthodes de la science ouverte s'avère être un processus relevant d'enjeux scientifiques, épistémologiques et techniques. Effectivement, la science ouverte s'affirme comme un nouveau paradigme dans le cycle de vie de la donnée, invitant à dépasser le cadre du projet pour se placer dans un mouvement d'ouverture et de partage des connaissances et techniques. En outre, les enjeux éthiques et juridiques qu'entraîne la science ouverte sont observables dans le

cadre de la gestion des données et des illustrations. Finalement, la mise en pratique de la science ouverte se traduit par l'adoption de choix techniques qui sont réalisés à toutes les échelles temporelles du projet, tant au niveau de l'encodage que la mise à disposition des données.

C'est cette mise à disposition et visualisation des notices du dictionnaire que nous avons pu esquisser au cours de cette partie, invitant à mettre en exergue les différentes échelles temporelles du projet. En effet, cette visualisation que nous proposons dans le cadre du prototype web implique de se questionner sur la pérennité des données mises à disposition par ce moyen et leur possible réutilisation. Il nous semble que les choix d'encodage et de développement dans langages et formats de données ouverts et interopérables sont des garants de cette (sur)vie de la donnée après le projet. De même, le recours aux outils et infrastructures développés et maintenus par HumaNum assurent une pérennité aux données produites. La temporalité du projet recoupe donc en partie le cycle de vie de la donnée qui se prolonge néanmoins au-delà de celui-ci.

Conclusion

Ce mémoire s'est attaché à mettre en évidence les différentes phases de traitement des données nativement numériques produites dans le cadre du projet ANR FermeGé. En premier lieu, la nécessité de développer un schéma d'encodage en XML/TEI appliqué aux notices historiques produites par l'équipe scientifique nous a invitée à questionner l'articulation entre attachement d'une grande précision sémantique et utilisation concrète des données. Effectivement, s'il est intellectuellement satisfaisant de proposer un modèle d'encodage extrêmement complet, encodant avec la granularité la plus fine les entités nommées, les dates, les mesures ou les monnaies, son application concrète dans la suite du projet n'est pas aisée à mettre à place. Effectivement, face à la quantité importante et toujours croissante de données textuelles produites, il était nécessaire de proposer un moyen d'encoder « automatiquement » ces mêmes notices. C'est pourquoi le schéma d'encodage TEI que nous avons rédigé sous forme d'un O.D.D se devait être la traduction technique des choix scientifiques. Ces orientations scientifiques nous ont été dictées par les besoins et objectifs de l'équipe d'historiens et d'historiennes cherchant à comprendre l'emprise spatiale de la Ferme générale sur la société d'Ancien Régime et sur une diversité remarquable de territoires. En conséquence, nous avons appliqué les décisions d'insister sur les entités nommées spatiales, à savoir les toponymes mentionnés dans les notices, ainsi que les organisations disposant d'un pouvoir sur un territoire identifié. C'est dans cette optique que l'O.D.D que nous avons d'abord rédigé, concernant individuellement aux notices, se présentait comme le fruit d'un dialogue entre technique et recherche scientifique.

La seconde phase de traitement des données a été l'encodage des notices. Pour ce faire, nous avons développé un processus d'encodage automatique reposant sur des expressions régulières par le biais du langage informatique Python. Cette phase de développement technique a été l'occasion de mettre en pratique les règles d'encodage théoriques précédemment évoquées. S'il est difficile de rendre concret et perceptible ce travail de développement, soulignons que ce fut l'occasion d'un travail par itération. Effectivement, après avoir encodé la structure des notices dites de « test », il nous a fallu « vingt fois sur le métier remettre notre ouvrage » afin d'obtenir un encodage des entités nommées satisfaisant. C'est dans cette phase de développement que nous avons pu mettre en application l'ontologie que l'équipe scientifique nous a transmise. Une fois que ce processus d'enco-

dage a atteint un résultat satisfaisant, nous avons envisager l'éditorialisation du contenu des notices au sein d'un prototype web.

La mise en place de ce prototype web a été pour nous l'occasion de réfléchir à l'inscription de la temporalité de notre stage dans un projet bien plus vaste. Effectivement, si notre prototype répond aux enjeux immédiats de mise à disposition du contenu scientifique des notices, nous avons fais le choix d'inclure d'autres fonctionnalités relatives à la continuité du projet, telles que l'indexation des entités nommées et des dates. Au grès du développement de ce prototype web et des discussions avec l'ensemble de l'équipe du projet, nous avons amplifié les fonctionnalités du site afin de positionner ce *Dictionnaire numérique de la Ferme générale* comme le point de confluence des différents axes du projet, ce qui n'est pas sans présenter certaines contraintes. Effectivement, dans la mesure où l'axe jumeau du dictionnaire, *l'Atlas de la Ferme générale*, n'évolue pas au même rythme, il nous a fallu anticiper et prévoir au mieux l'interopérabilité des données textuelles et géographiques. En tout état de cause, l'insertion dans les principes et méthodes de la science ouverte s'avère être un moyen pertinent de pallier certains de ces enjeux, tout en assurant une certaine pérennité au projet.

Finalement, ce stage et la rédaction de ce mémoire ont été pour nous l'occasion d'amener une réflexion critique sur notre positionnement et parcours dans l'ingénierie en humanités numériques associée au monde de la recherche en histoire. En somme, en quoi le travail que nous avons effectué nous permet de nous projeter mentalement et professionnellement dans le domaine des humanités numériques et du traitement des données scientifiques ? Nous proposons de répondre à cette ultime problématique par un détour bibliographique. Alors que notre stage débutait, paraissait en librairie le troisième tome de la trilogie consacré à l'ordre matériel du savoir, par l'historienne des sciences, et directrice de recherche émérite au CNRS, Françoise Waquet. Ce troisième ouvrage intitulé *Dans les coulisses de la science. Techniciens, petites mains et autres travailleurs invisibles*³⁹ propose une analyse historique de la réalité des conditions de travail de ceux qui agissent « dans les sous-sols de la tour d'ivoire⁴⁰ », à savoir les garçons de laboratoire des grands biologistes, les techniciens assurant le bon déroulé des recherches dans les domaines de la physique et de la chimie, les épouses de chercheurs relisant les épreuves et bien évidemment (et plus tardivement) les ingénieurs d'études et de recherche appuyant les programmes de recherche. Au cours de son étude abordant les évolutions séculaires (entre la fin du XVIII^e siècle et le début du XXI^e) et la position de ces « petites mains de la recherche⁴¹ », l'historienne souligne l'inflation du nombre des ingénieurs et techniciens

39. Françoise Waquet, *Dans les coulisses de la science : techniciens, petites mains et autres travailleurs invisibles*, Paris, France, 2022.

40. *Ibid.*, p.13

41. *Ibid.*, p.21.

qu’implique « l’institutionnalisation de la science⁴² » avec la création du CNRS. Cet accroissement du nombre d’ingénieurs et leur montée en qualification universitaire comme technique entraîne un questionnement sur leur place dans l’élaboration de la connaissance et du processus de recherche scientifique. Si leur tâche relève a priori de l’exécution⁴³ des directives scientifiques, leur travail quotidien participe à différents niveaux à la mise en oeuvre des réflexions scientifiques⁴⁴. Il en ressort d’après cette étude que les ingénieurs, tout particulièrement en humanités numériques, se positionnent à l’interface de la recherche et de l’exécution technique. Cette situation semble se renforcer par l’enjeu de pouvoir former un véritable réseau-métier à part entière, qui existe déjà pour d’autres emplois techniques au sein du CNRS⁴⁵. Ces analyses rassemblées par le regard expert de la chercheuse ne pouvaient qu’entrer en résonance avec notre ressenti personnel à l’issu du stage. Effectivement, nous avons pu participer à notre échelle à un projet où les ingénieurs sont des intermédiaires entre les besoins scientifiques du projet et les enjeux techniques de gestion des données. Ce dialogue fécond nourrit, d’après les propos de Marie-Laure Legay, coordinatrice du projet, la réflexion scientifique et « motive » les historiens dans la rédaction des notices. Ainsi, d’un point de vue personnel, cette première expérience de l’ingénierie en humanités numériques a été pour nous une confirmation d’un choix d’orientation et l’ouverture vers un nouvel horizon professionnel.

42. *Ibid.*, p. 59.

43. *Ibid.*, voir chapitre 4 « Exécuter »

44. *Ibid.*, voir chapitre 5 « Oeuvrer »

45. *Ibid.*, p. 77

Acronymes

AD Archives Départementales.

AM Archives Municipales.

AN Archives Nationales.

ANR Agence Nationale de la Recherche.

BNF Bibliothèque nationale de France.

CEPRISCA CEntre de droit PRivé et de Sciences Criminelles d'Amiens (UR 3911).

CHAD Centre d'Histoire et d'Anthropologie du Droit (UR 4417).

CMS Content Managing System.

CNRS Centre National de la Recherche Scientifique.

CRESAT Centre de Recherches sur les Economies, les Sociétés, les Arts et les Techniques (UR 3436).

CSS Cascading Style Sheets.

CSV *Comma-Separated values*.

CTHS Comité des Travaux Historiques et Scientifiques.

DOI Digital Object Identifier.

DRES Droit, Religion, Economie et Société (UMR 7354).

DTD Document Type Definition.

ENC École nationale des chartes.

GIP Groupement d'Intérêt public.

HNOMAD Humanités Numériques, Outils, Méthodes et Analyse de Données.

HTML HyperText Markup Language.

HTR Handwritten Text Recognition.

IFRESI Institut Fédératif de Recherche - Économie et Sociétés Industrielles.

IIIF International Image Interoperability Framework.

IRHiS Institut de Recherches Historiques du Septentrion (UMR 8529).

JSON JavaScript Object Notation.

LMF Lexical Markup Framework.

MESHS Maison Européenne des Sciences de l'Homme et de la Société.

MSH Maison des Sciences de l'Homme.

OCR Optical Character Recognition.

ODD One Document Does it all.

ORM Object-Relation Mapping.

PHP Personnal Home Page.

RDF Resource Description Framework.

RNG Regular Language for XML Next Generation.

RNMSH Réseau National des Maisons des Sciences de l'Homme.

SHS Sciences Humaines et Sociales.

SIAF Service Interministériel des Archives de France.

SQL Structured Query Language.

TAL Traitement Automatique des Langues.

TEI Text Encoding Initiative.

TGIR Très Grande Infrastructure de Recherche.

UAR Unité d'Appui à la Recherche.

UDL Université de Lille..

ULCO Université du Littoral Côte d'Opale..

UMR Unité Mixte de Recherche.

UPJV Université de Picardie Jules Verne.

UR Unité de Recherche.

VIAF Virtual International Authority File.

XML eXtended Markup Language.

XSLT eXtensible Stylesheet Language Transformation.

Glossaire

API Application Protocol Interface, système informatique prévu pour la communication de données ou de fonctions de manière automatisée à des services tiers..

Entités nommées expression linguistique référentielle faisant référence à un toponyme, un anthroponyme ou une organisation spécifique..

Espace de noms ou *namespace*, soit un lieu abstrait conçu pour accueillir des ensembles de termes appartenant à un même répertoire. En informatique, cela permet de lever l'ambiguïté sur l'utilisation de certains termes ou certaines données..

Expressions Régulières (Regular Expressions) motifs exprimé sous forme d'une chaîne de caractères qui décrit un ensemble de chaînes de caractères possibles..

Granularité caractérise la précision d'un modèle de données et sa capacité à décrire en finesse les entités encodées..

Ontologie en informatique, caractérise un modèle de données associé à un domaine de connaissance particulier.

Parsing en informatique, segmentation d'un texte suivant des paramètres prédéfinis, généralement en vue d'une analyse syntaxique ou de l'exploitation d'une ressource textuelle..

Web de données modèle de donnée assurant leur échange et leur interopérabilité en se basant sur le *Resource Description Framework* dans le cadre de l'architecture web..

Workflow Processus ou flux opérationnel de traitement des données divisé en différentes tâches complémentaires..

Annexes

A. Dépôt Gitlab

Les annexes de ce présent mémoire, incluant les livrables du stage, sont disponibles dans le dépôt Gitlab trouvable à l'adresse suivante :<https://gitlab.huma-num.fr/vdecreaene/DicoNumFermeGe>. Nous les avons en partie reproduites ici afin d'accompagner et compléter notre travail.

.1 Arborescence



FIGURE 7 – Arborescence du dépôt Gitlab

.2 Présentation des dossiers

Le dépôt Gitlab est contenu dans un dossier général intitulé « DicoNumFermeGe ». Il s'agit de la racine du dépôt qui contient les éléments structurants suivants :

— « **Scripts_encodage_FermeGe** » :

Dossier composé des scripts Python et des notices automatiquement encodées en XML/TEI. Il s'agit de la chaîne de traitement de la donnée initiale qui prend en entrée une notice rédigée dans un document texte (« txt ») et produit en sortie une notice encodée (« output ») en accord avec le modèle conceptuel et l'O.D.D réalisés par l'équipe. Le premier script (« script_encodage_v1.py ») produit des notices individuelles dans l'optique du dépôt pérenne sur Nakala, puis sont ensuite fusionnées (« script_encodage_teiCorpus.py ») au sein d'un corpus intermédiaire (« corpus_intermediaire.xml »), et enfin mise en forme au sein du corpus final (« corpus.xml »). Ce fichier « corpus.xml » contient donc l'ensemble de l'échantillon des notices (29 entrées de la lettre A) et sert de base de fonctionnement et de visualisation au sein de l'application web.

— « **app** » :

Contient l'ensemble des dossiers et fichiers propres au fonctionnement de l'application web.

Le dossier « modeles » permet de développer un ORM faisant le lien avec la base de données relationnelle (« Base_de_donnees_Ferme_Ge.sql ») par le biais des différentes classes déclarées. Au sein des modeles, le fichier « donnees.py » contient les classes et le script d'initialisation de la base de données. Le fichier « utilisateurs.py » a été ajouté en prévision de l'implémentation des comptes utilisateurs gérés grâce aux classes correspondantes.

Le dossier « routes » contient, au sein du fichier « generic.py », l'ensemble des fonctions permettant la création et visualisations des différentes routes (ou pages) de l'application. Nous avons eu recours ici au micro-framework Flask, développé en Python, permettant d'assurer la structuration du site, la génération automatique des pages en fonction du nombre de notices, ect. Un second dossier pourra par la suite être ajouté par l'équipe de développement pour implémenter une API afin de mettre à disposition les données produites par le projet et assurer un ancrage dans les principes de la science ouverte.

Le dossier « static » contient l'ensemble des sous-dossiers relevant des éléments « statiques » de l'application, à savoir :

1. la police de caractère (sous-dossier « fonts » contenant les fichiers nécessaires).
2. les images (« images » : illustrations, bannières et logos des institutions).
3. les librairies CSS et JavaScript (« css » et « js ») associées au framework

Bootstrap que nous avons utilisé pour l'aspect visuel.

4. le corpus XML/TEI des notices issues de la chaîne de traitement (sous-dossier « xml » et fichier « corpus.xml »)
5. les feuilles de transformation XSLT (« xsl ») permettant le passage des notices du format XML/TEI au format web HTML. Le fichier « affichage_notice.xsl » est une feuille de transformation permettant la visualisation des notices suivant le modèle établis par l'équipe, à savoir la segmentation entre le titre, le corps du texte, les références bibliographiques et archivistiques et les notices liées. Les fichiers « chronologie.xsl », « index_orga.xsl » et « index_place.xsl » permettent l'indexation des entités nommées suivant l'ontologie établie au sein du projet.

Le dossier « templates » contient les différents documents HTML qui sont utilisés par les fonctions du fichier « generic.py » pour générer les pages. Les templates accueillent donc les données et permettent une structuration des pages web grâce au format HTML. Le sous-dossier « pages » contient l'ensemble des templates HTML de base, qui peuvent être eux-même insérés dans des conteneurs affichant les onglets et barres de navigation (voir sous-dossier « partials »). Le fichier « constantes.py » contient les valeurs constantes de certaines variables, telles que le nombre de pages ou de résultats à afficher, que nous appelons ponctuellement au sein de nos fonctions « generic ».

— « **ODD_schemas** » :

Contient les schémas d'encodage et la documentation associée permettant de valider la conformité des notices. Ces ODD découlent du processus du modèle conceptuel développé au sein de l'équipe du projet. Le fichier « ODD_v2.xml » contient l'ODD appliqué aux notices individuelles (issues donc du script « script_encodage_v1.py ») rédigé dans la « syntaxe ODD » au format XML/TEI. La première partie de ce fichier contient la documentation en prose et la seconde partie le schéma d'encodage. Cet ODD a été transformé au format RNG par un processus Saxon PE au sein du logiciel Oxygen afin d'être fonctionnel. Le même processus a été effectué pour l'ODD appliqué au corpus (« ODD_teicorpus.xml » et « ODD_teicorpus.rng »). Ces mêmes fichiers au format HTML permettent d'obtenir une visualisation web des différentes ODD une fois téléchargés et lus par un navigateur web.

— « **SortieTex_PDF et SortieXML_TeX** » :

Ces deux dossiers comprennent l'ensemble des transformations des notices (initialement au format XML) vers le format TeX, puis vers le format PDF afin de rendre possible le téléchargement du contenu scientifique des pages web. Les no-

tices au format TeX ont été automatiquement générées par le biais d'une feuille de transformation XSLT (grâce au paramètre de sortie "text", il nous a été possible d'injecter des commandes LaTeX dans la transformation). Les sorties au format TeX ont ensuite été transformées par un compilateur LaTeX en fichier PDF, que nous appelons par le biais d'une fonction dans notre fichier generic.py

B. Schématisation et documentation

.3 ODD des notices individuelles

L'ODD des notices individuelles se trouve à l'adresse suivante : https://gitlab.huma-num.fr/vdecreaene/DicoNumFermeGe/-/blob/main/ODD_schemas/ODD_v2.xml. Une sortie HTML permettant une lecture de la documentation dans un navigateur web est disponible ici : https://gitlab.huma-num.fr/vdecreaene/DicoNumFermeGe/-/blob/main/ODD_schemas/ODD_v2.html (téléchargez puis ouvrez le fichier ; une page web s'exécutera).

.4 ODD du corpus de notices

L'ODD appliqué au corpus se trouve à l'adresse suivante : https://gitlab.huma-num.fr/vdecreaene/DicoNumFermeGe/-/blob/main/ODD_schemas/ODD_teiCorpus.xml. Une sortie HTML permettant une lecture de la documentation dans un navigateur web est disponible ici : https://gitlab.huma-num.fr/vdecreaene/DicoNumFermeGe/-/blob/main/ODD_schemas/ODD_teiCorpus.html (téléchargez puis ouvrez le fichier ; une page web s'exécutera).

.5 Ontologie et documentation

Le document de travail ayant donné lieu à l'ontologie d'encodage des entités nommées est constitué d'un tableau *framacalc* ne pouvant être reproduit dans son intégralité ici. Il est cependant librement accessible à l'adresse suivante : <https://lite.framacalc.org/fermeege-ontologie-des-entites-nommees-9udi.html>

C. Scripts Python

L'ensemble des scripts Python commentés sont disponibles dans le dossier du dépôt Gitlab présent à l'adresse suivante : https://gitlab.huma-num.fr/vdecreaene/DicoNumFermeGe/-/tree/main/Scripts_encodage_FermeGe/scripts. Ci-dessous nous ne reproduisons dans cette section que les extraits de code que nous mentionnons au cours du chapitre 5.

.6 Script d'encodage des notices individuelles

```
def build_tei_container():
    container = f"""<?xml-model href="file:/F:/Stage%20MESHS%20DE%20CRAENE%20Valentin/TEI%20XSLT/out/ODD_v2.rng" type="application/xml"
<?xml-model href="file:/F:/Stage%20MESHS%20DE%20CRAENE%20Valentin/TEI%20XSLT/out/ODD_v2.rng" type="application/xml" schematypens="http://www.tei-c.org/ns/1.0"></TEI>"""
    parsed_container = BeautifulSoup(container, "xml")
    parsed_teи_header = BeautifulSoup(build_teи_header(), "xml").extract()
    parsed_container.TEI.append(parsed_teи_header)
    parsed_container.TEI.append(parsed_container.new_tag("text_blob"))
    return parsed_container
```

FIGURE 8 – Script Python : fonction de création du conteneur TEI.

```
def add_notice_title(text):
    text = re.findall(r'^[a-zA-ZÀÁÇÈÉËÍÏÒÓÙÜÝÀÁÇÈÉËÍÏÒÓÙÜÝ]([àáçèéëíïòóùüý])', str(text), re.MULTILINE)
    text = ''.join(text)
    text = '<form type="lemma"><orth>ph</orth></form>'.replace('ph', str(text))
    tree = BeautifulSoup(str(text), 'xml')
    return tree
```

FIGURE 9 – Script Python : fonction d'encodage des titres.

```
def fix_idno_tags(text):
    text = text.replace("&lt;idno;", "<idno").replace("&lt;/idno&gt;", "</idno>")
    return text.replace("&gt;", ">").replace("&lt;", '<')
```

FIGURE 10 – Script Python : fonction de correction des balises <idno>.

```
def add_scientific_references(text):
    text = re.findall(r'\n[.]+\n$', str(text), re.MULTILINE)
    text = ''.join(text)
    text = '<bibl type="references">ph</bibl>'.replace('ph', str(text))
    tree = BeautifulSoup(str(text), 'xml')
    return tree
```

FIGURE 11 – Script Python : fonction d’encodage de la bibliographie (niveau 1).

```
def add_scientific_references_single(text):
    list = []
    text = re.findall(r'^(.*\n)', str(text), re.MULTILINE)
    text = ' '.join(text)
    text = text.split(';')
    for i in text:
        text = list.append('<bibl>ph</bibl>'.replace('ph', i+";"))
    return ''.join(list)
```

FIGURE 12 – Script Python : fonction d’encodage de la bibliographie (niveau 2).

```
def find_sources(complete_tree):
    complete_tree = main('txt', 'test2')
    complete_tree = BeautifulSoup(complete_tree, 'xml')
    regex_sources = re.findall(r'(Arrêt.*?;| Arrest.*?; | Lettre.*?;| Ordonnance.*?;| Carte.*?;| Vue.*?;| Edit.*?;| Édit.*?;| Déclarat')
    regex_sources = ''.join(regex_sources)
    regex_idno = re.compile(r'(AM .*;|AN,? ?.?;|An .*;|AD,? ?.?;|BNF .*;|(?<!,)( (G\d.*?;)| \dC.*?;))')

    for tag in complete_tree.bibl.find_all("bibl"):
        if tag.string:
            tag.string.replace_with(re.sub(regex_idno, r'<idno type="ArchivalIdentifier">\1</idno>', tag.string, 0))
        if re.search(r'(Arrêt.*?;| Arrest.*?; | Lettre.*?;| Ordonnance.*?;| Carte.*?;| Vue.*?;| Edit.*?;| Édit.*?;| Déclaration.*?;| A')
            tag.attrs['type'] = 'sources'
    return complete_tree.prettify()
```

FIGURE 13 – Script Python : fonction d’encodage et de différenciation des sources.

7 Script d’encodage du corpus TEI

```
def build_final_tree():
    filenames = glob.glob('scripts/output/*.xml')
    outfilename = 'corpus_intermediaire.xml'

    with open(outfilename, 'wb') as outfile:
        for filename in glob.glob('scripts/output/*.xml'):
            if filename == outfilename:
                # don't want to copy the output into the output
                continue
            with open(filename, 'rb') as readfile:
                shutil.copyfileobj(readfile, outfile)
```

FIGURE 14 – Script Python : extrait de la fonction de fusion des notices individuelles.

```

with open('corpus_intermediaire.xml', 'r', encoding='utf-8') as file:
    file = file.read()
    file = file[:38] + "<teiCorpus xmlns='http://www.tei-c.org/ns/1.0' version='1.0'>" + file[39:] + '</teiCorpus>'
    file = BeautifulSoup(file, 'xml')
    file = str(file).replace('<?xml version="1.0" encoding="utf-8"?>', '')
    file = str(file).replace('<?xml-model href="file:/F:/Stage%20MESHS%20DE%20CRAENE%20Valentin/TEI%20XSLT/out.xsl?mode=match'"?>', '')
    file = '<?xml version="1.0" encoding="utf-8"?>' + file

```

FIGURE 15 – Script Python : extrait de la fonction de création d'un corpus intermédiaire.

```

complete_tree = BeautifulSoup(file, 'xml')
complete_tree = add_n_counters_to_tei(complete_tree)
complete_tree = add_n_counters_to_entry(complete_tree)
complete_tree = del_namespace(complete_tree)
header = build_teiCorpus_header()
header = BeautifulSoup(header, 'xml')
complete_tree.teiCorpus.insert(0, header)

```

FIGURE 16 – Script Python : extrait de la fonction de mise en conformité du corpus.

```

with open('corpus.xml', "w", encoding="utf8") as fh:
    fh.write(str(complete_tree))

```

FIGURE 17 – Script Python : extrait de la fonction d'enregistrement du corpus.

D. Guide d'installation de l'application web

« DicoNumFermeGé »

Nota bene : commandes à exécuter dans un terminal type shell Unix (Linux ou macOS). L'usage d'une machine virtuelle pour les systèmes d'exploitation Windows est fortement recommandé.

- Cloner le dossier :

```
git clone https://gitlab.huma-num.fr/vdecreaene/DicoNumFermeGe.git
```

- Télécharger un package d'environnement virtuel :

```
sudo apt-get install python3 libfreetype6-dev python3-pip python3-virtualenv
```

- Vérifier que la version de Python est supérieure à 3.7 :

```
python --version
```

- Installer l'environnement virtuel :

```
virtualenv ~/.DicoNumFermeGe -p python3
```

- Activer l'environnement :

```
source ~/.DicoNumFermeGe/bin/activate
```

- Lancer la commande (télécharger les modules et dépendances nécessaires) :

```
pip install -r requirements.txt
```

- Lancer l'application :

```
python3 run.py
```

- Aller sur <http://127.0.0.1:5000/> ou cliquer sur ce même lien dans votre IDE

142D. GUIDE D'INSTALLATION DE L'APPLICATION WEB « DICONUMFERMEGÉ »

E. Application web

« DicoNumFermeGé »

Nous reproduisons ci-dessous par le biais de captures d'écrans les pages de l'application web que nous commentons dans le corps du mémoire.

Dictionnaire Numérique de la Ferme générale

Le projet FermeGé vise à étudier l'impact d'une organisation fiscale (1664-1794), discriminante mais rationnelle, sur les territoires et les sociétés de la France moderne. Il cerne les dynamiques de fonctionnement d'une institution ancrée dans une culture d'inégalité, mais tout autant dans une culture administrative éclairée visant l'efficacité.

Véritable « Etat », omniprésente sur des territoires très différenciés pour collecter près de 50 % des revenus ordinaires de la monarchie, dotée de moyens exceptionnels de coercition, mais capable de transactions, la FG a localement renforcé ou affaibli le sentiment d'injustice à l'interface avec les sociétés plurielles sur lesquelles elle agit. La confrontation entre une logique gestionnaire éclairée par une science et un droit administratif nouveaux d'une part, et des identités géographiques et sociales plurielles générée des réactions qui se déclinent en pratiques et discours pluri-sémantiques sur l'inégalité, allant jusqu'à la radicalisation violente, mais également en pratiques de conciliation caractéristiques de l'arbitrage administratif et gestionnaire. Le projet s'appuie sur une collaboration de quatre laboratoires Histoire/Géographie/Histoire du droit en prenant acte des renouvellements heuristiques de chaque discipline. Il vise à restituer des connaissances inédites sur cette organisation fiscale et des analyses sur l'interface inégalitaire impôts/territoires/sociétés qui ne se limitent pas au paradigme d'une organisation purement coercitive. En effet, l'impôt constitue dès l'époque moderne un outil de réduction des inégalités grâce à la rationalité administrative, fonction reprise par l'Etat contemporain.

Au-delà, l'originalité du projet réside encore dans le questionnement d'un binôme notionnel « inégalité/rationalité » que nous élaborons à partir d'un modèle de gestion de l'inégalité qui contribua à l'émergence du droit administratif français et a été exporté à l'étranger. Nous émettons l'hypothèse que ce binôme est opérationnel pour étudier tout type d'organisation fiscale agissant globalement sur un territoire.

Recherche rapide

Rechercher

Accueil A propos ▾ Découvrir le projet ▾ Rechercher Index et chronologie ▾

Quai Matignon x

Acquit à caution

Adjudicataires

Agents (voir Tabac)

Agent

Agent général des fermes au Canada (voir Canada)

Aides

Aiguës-Mortes

Alâtre Julien (voir Adjudication)

Allège

Alsace

Alun (droits sur)

Amérique (voir Colonies)

Amidon (Droits sur)

Anjou

Annuel (droit)

Arc-et-Senans

Arles

Armée (voir Soldat)

FIGURE 18 – Application web : page d'accueil

Axe 1 : Dictionnaire numérique de la Ferme générale : Index des notices			
Identifiant	Titre	Auteur	Lien vers la notice
1	Acquit à caution	Marie-Laure Legay	Acquit à caution
2	Adjudicataire	Marie-Laure Legay	Adjudicataire
3	Agenais (voir Tabac)	Marie-Laure Legay	Agenais (voir Tabac)
4	Agent	Thomas Bouillu	Agent
5	Agent général des fermes au Canada (voir Canada)	Marie-Laure Legay	Agent général des fermes au Canada (voir Canada)
6	Aides	Marie-Laure Legay	Aides
7	Aiguës-Mortes	Marie-Laure Legay	Aiguës-Mortes
8	Alatierre Julien (voir Adjudication)	Marie-Laure Legay	Alatierre Julien (voir Adjudication)
9	Allège	Marie-Laure Legay	Allège
10	Alsace	Marie-Laure Legay	Alsace
11	Alun (droits sur)	Marie-Laure Legay	Alun (droits sur)
12	Amérique (voir Colonies)	Marie-Laure Legay	Amérique (voir Colonies)

FIGURE 19 – Application web : liste des notices incorporées (axe 1).

The screenshot shows a web page with a dark red header bar containing navigation links: 'A propos', 'Découvrir le projet', 'Rechercher', 'Index et chronologie', and 'Rechercher'. A search input field with placeholder 'Rechercher' is also present. The main content area has a light gray background.

Axe 4 : Un ancrage dans les principes de la science ouverte

En cours de développement

Documentation du projet

O.D.D et schémas d'encodage du projet

Documentation des schémas d'encodage du projet (O.D.D au format .xml)

Présentation : Cette section contient la documentation des deux schémas d'encodage appliqués au projet, sous forme d'O.D.D comprenant d'une part une présentation en prose des choix d'encodage et un schéma respectant les principes de l'O.D.D, destiné à être transformé au sein d'un logiciel d'encodage de type Oxygen afin d'obtenir une sortie .rng. Le premier O.D.D valide l'encodage automatique des notices individuelles alors que le second schématise le corpus TEI au sein de cette présente application.

Lien vers le : Gitlab

Télécharger l'ODD des notices au format XML

Télécharger l'ODD du corpus au format XML

Télécharger l'ODD des notices au format HTML

Télécharger l'ODD du corpus au format HTML

Titre de la documentation

Description succincte de la documentation

Titre de la documentation

Description succincte de la documentation

FIGURE 20 – Application web : science ouverte et documentation (axe 4).

A propos ▾ Découvrir le projet ▾ Rechercher Index et chronologie ▾

Recherche rapide Rechercher

Index des lieux d'habitations.

☰ Liste des notices publiées

Dénomination	Lien vers la notice
ports de Dunkerque	Acquérir à caution
ville Rouen	Aides
Ville de Camargue	Aigues-Mortes
ville de Strasbourg	Alsace
ports de Marseille	Alun_droits_sur
villes d'Angers	Anjou
villes de Mayenne	Anjou
ville de Blois	Annuel_droit
ville de Grandville	Annuel_droit
village nommé Arc	Arc-et-Senans
village de Senans	Arc-et-Senans
ville d'Arles	Arles

FIGURE 21 – Application web : index topographique.

 A propos ▾ Découvrir le projet ▾ Rechercher Index et chronologie ▾

Recherche rapide Rechercher

Liste des images, graphiques et reproductions d'archives :

#	Titre de la notice	Auteur	Titre du document	Source / référence scientifique	Mentions légales
1	Adjudicataire	Marie-Laure Legay	Plomb de scellé de la Ferme générale, 1732, bureau de Caen	Référence à voir avec M. L. Legay : Antoine Sabatier, <i>Sigillographie historique des administrations fiscales, communautés ouvrières et institutions diverses ayant employé des sceaux de plomb (XIV-XVIII siècles)</i> ; Plombs historiques de la Saône et de la Seine, Paris, H. Champion, 1912	A déterminer. Autorisation de reproduction des archives par la loi du 28 décembre 2015 (loi Valter)
2	Aides	Marie-Laure Legay	Recette des droits d'aides au sein de la Direction de Lyon (1788), en livre tournois	Tableau produit en interne	Demande d'autorisation de reproduction à l'auteur
3	Aides	Marie-Laure Legay	Nombre de procès-verbaux de fraude aux droits des aides	Tableau produit en interne	Demande d'autorisation de reproduction à l'auteur
4	Alsace	Marie-Laure Legay	Produits de la régie des aides par direction en Alsace (AN, G1 131, 1731)	Tableau produit en interne à partir de AN, G1 131, 1781	Demande d'autorisation de reproduction à l'auteur pour le tableau. Autorisation de reproduction des archives par la loi du 28 décembre 2015 (loi Valter)

FIGURE 22 – Application web : mentions légales.

F. Workflow final

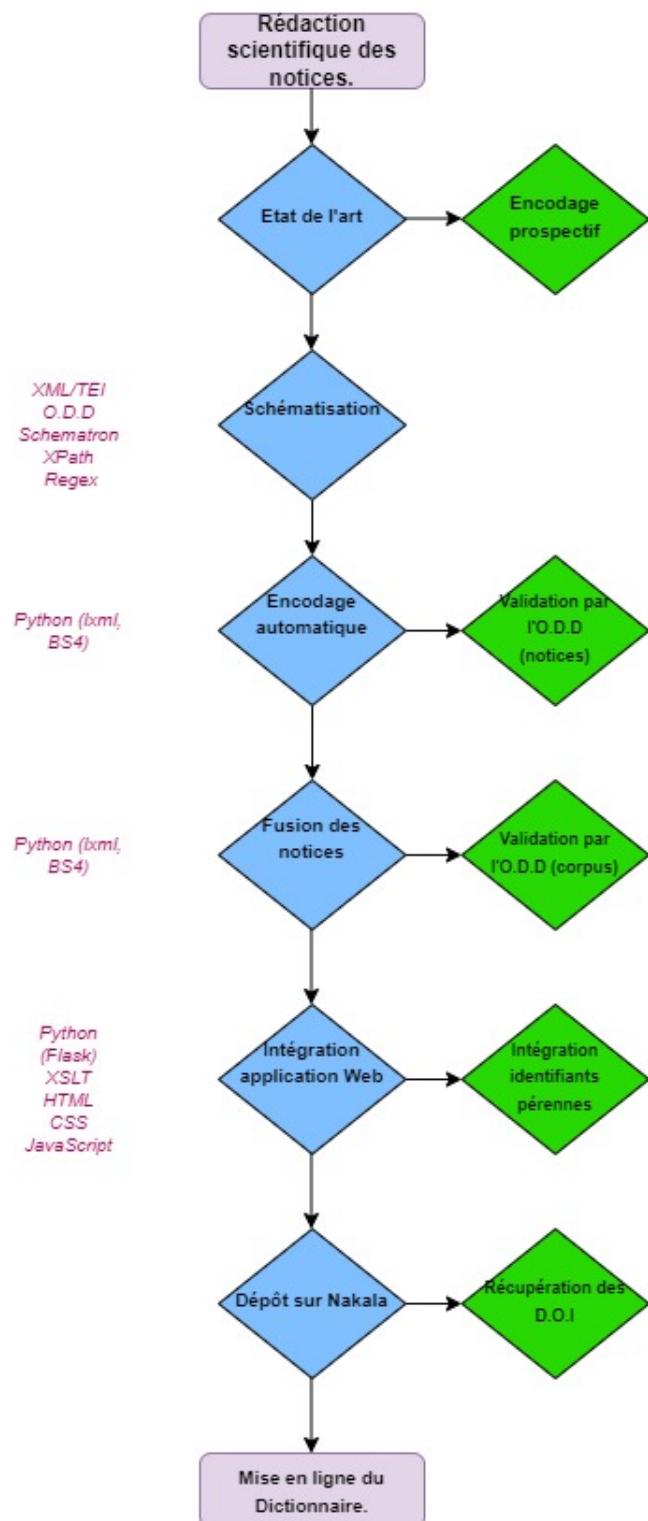


FIGURE 23 – Workflow final.

G. Arborescences XML/TEI

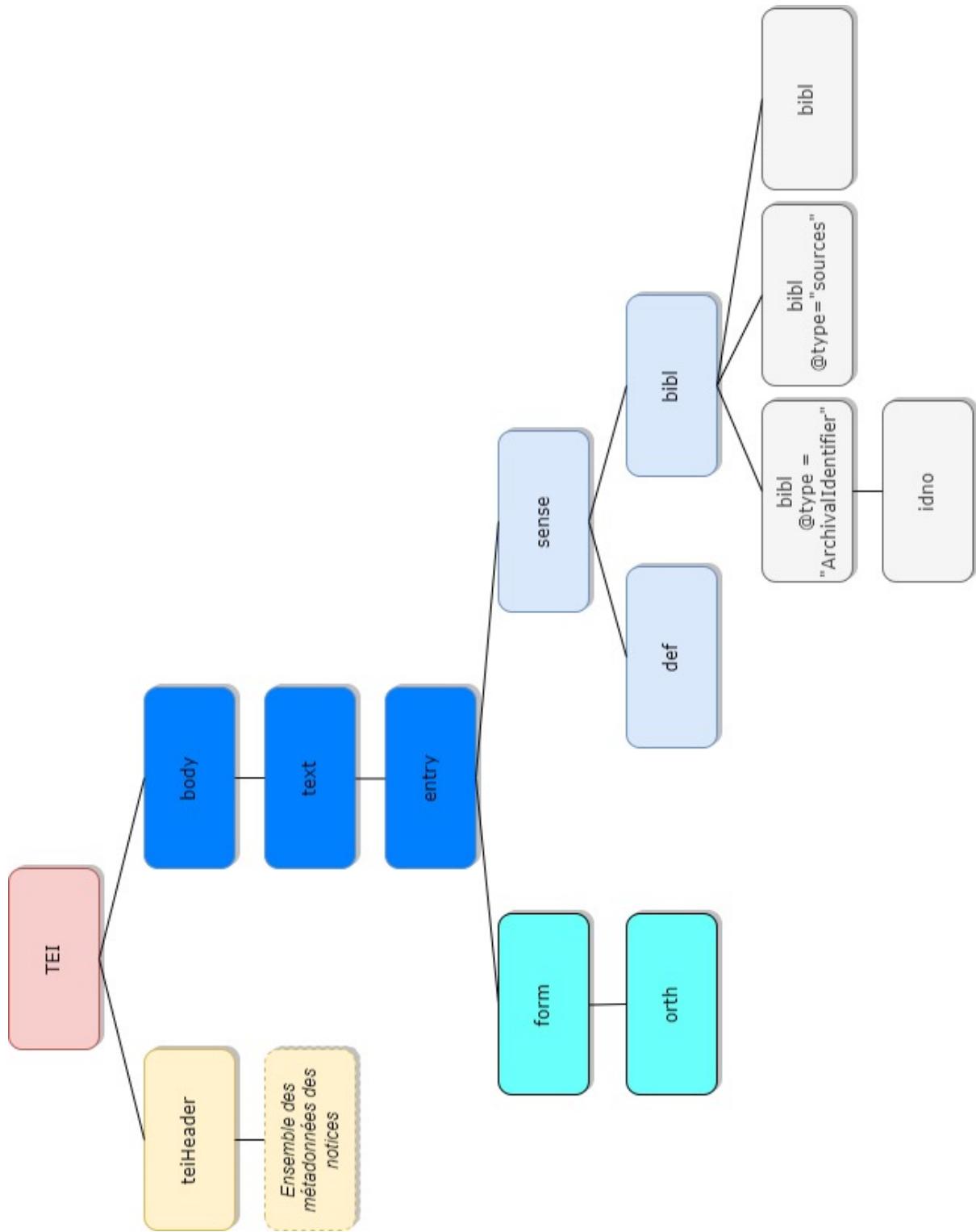


FIGURE 24 – Arborescence XML/TEI des notices.

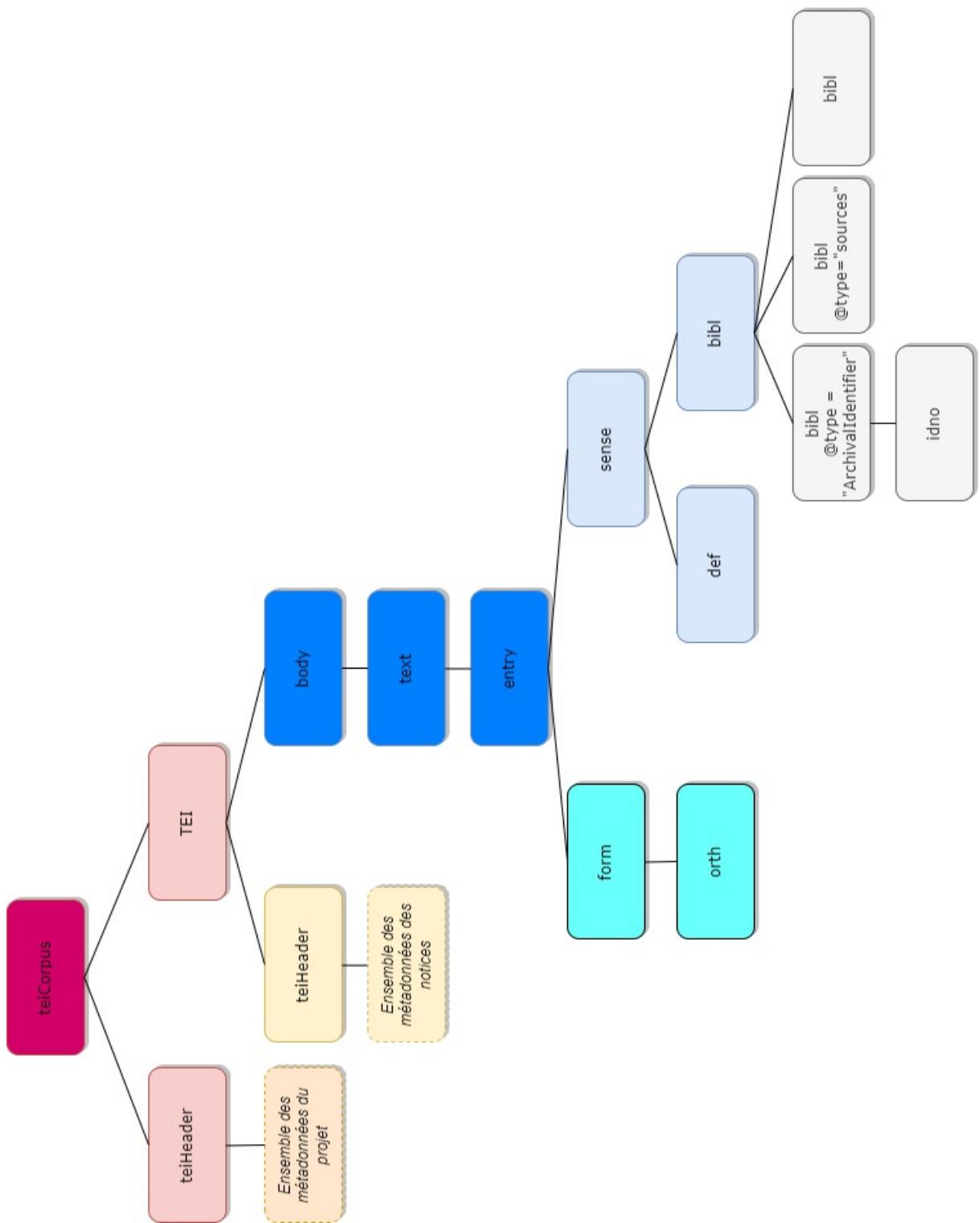


FIGURE 25 – Arborescence XML/TEI des notices.

Table des figures

1.1	Carte des laboratoires affiliés à la MESHS, 2022.	12
1.2	Diagramme des outils d’HumaNum utilisés dans le <i>workflow</i> du projet.	13
2.1	<i>Workflow</i> : état de l’art de l’encodage des dictionnaires.	16
2.2	Modèle d’encodage minimal en TEI des notices de dictionnaires	18
2.3	Notice « Ferme générale » dans le document de travail du Dictionnaire.	18
2.4	Encodage XML/TEI de la notice PUTAGIUM du Dictionnaire de Ducange en ligne.	20
2.5	Encodage de la notice ”neo” tirée du projet Nénufar	21
2.6	Encodage de la notice ”gagnant” tirée du projet Basnage.	22
2.7	Exemple d’encodage automatisé d’une entrée avec Grobid-dictionaries	24
3.1	<i>Workflow</i> : phase d’encodage manuel et prospectif	28
3.2	Schéma du « modèle conceptuel » du Dictionnaire de la Ferme générale.	29
3.3	Schéma du « modèle logique » du <i>Dictionnaire de la Ferme générale</i>	30
3.4	Encodage XML/TEI minimal de la structure d’une notice.	31
3.5	Extrait de l’encodage du fichier teiCorpus du projet ThEMA.	32
3.6	Diagramme la reconnaissance d’entités nommées dans les différentes approches des humanités numériques (TAL/NLP)	36
3.7	Capture d’écran : proposition d’encodage des anthroponyme sous forme d’entités unitaires.	36
3.8	Capture d’écran : proposition d’encodage des anthroponymes avec une granularité fine.	37
3.9	Capture d’écran : proposition d’encodage des entités morales.	37
3.10	Capture d’écran : proposition d’encodage de ”persName” avec insertion du référentiel VIAF.	37
3.11	Capture d’écran : proposition d’encodage de ”orgName”	37
4.1	<i>Workflow</i> : Schématisation et conceptualisation de l’O.D.D	45
4.2	Schéma : modèle logique en arborescence XML/TEI des notices.	46
4.3	Capture d’écran : déclaration des modules additionnels « namesdates », « dictionnaries » et « figures ».	47

4.4	Capture d'écran : schéma O.D.D et <i>Schematron</i> , extrait du schéma appliquée à la balise <entry>	47
4.5	Capture d'écran : règles de l'O.D.D portant sur l'ordre des balises au sein de la balise <sense>.	48
4.6	Capture d'écran : déclaration des éléments autorisés du module « header »	49
4.7	Capture d'écran : indexation matière RAMEAU des notices du dictionnaire.	50
4.8	Capture d'écran : exemple d'encodage des sources et de la bibliographie de la balise « amidon ».	52
4.9	Capture d'écran : extrait de l'O.D.D appliqué à la balise <bibl>	53
4.10	Schéma : <i>workflow</i> et interaction des deux O.D.D	55
4.11	Schéma : arborescence XML/TEI du corpus TEI pour l'application web.	56
4.12	Capture d'écran : ajout du teiCorpus dans le fichier .rng à la main	57
4.13	Capture d'écran : extrait de l'ODD pour les organisations et institutions (orgName).	61
4.14	Capture d'écran : extrait de l'ODD pour les toponymes (placeName).	62
4.15	Capture d'écran : extrait de l'O.D.D concernant les dates.	63
5.1	<i>Workflow</i> : phase d'encodage automatisé des notices.	66
5.2	Tableau de présentation des <i>parsers</i> en Python, extrait de la documentation BS4.	67
5.3	Diagramme : fonctionnement algorithmique de l'encodage de la bibliographie et des sources.	71
6.1	<i>Workflow</i> : phase de développement de l'application web	82
6.2	Fonctionnement des <i>templates</i> avec Flask.	87
6.3	Schéma fonctionnement de l'application web Flask.	90
6.4	Carte CRESAT - Daniel Fischer, « Le « mille feuille» territorial alsacien à la veille de la Révolution : souveraineté française, seigneuries étrangères », in <i>Atlas historique d'Alsace</i> , www.atlas.historique.alsace.uha.fr , Université de Haute Alsace, 2011	95
7.1	<i>Workflow</i> : phase d'ouverture et mise à disposition des données.	100
7.2	Schéma : cycle de vie de la donnée et science ouverte dans le projet FermeGé.104	104
7.3	Tableau : synthèse des enjeux juridiques par type de documents.	109
7.4	Tableau comparatif des formats de données au sein d'une API.	112
7.5	Schéma illustrant l'insertion des <i>datapapers</i> dans l'écosystème des publications académiques.	114
7.6	Tableau comparatif des éléments relevant d'un dictionnaire et d'une base de données.	116
7	Arborescence du dépôt Gitlab	130

8	Script Python : fonction de création du conteneur TEI.	137
9	Script Python : fonction d'encodage des titres.	137
10	Script Python : fonction de correction des balises <idno>.	137
11	Script Python : fonction d'encodage de la bibliographie (niveau 1).	138
12	Script Python : fonction d'encodage de la bibliographie (niveau 2).	138
13	Script Python : fonction d'encodage et de différenciation des sources.	138
14	Script Python : extrait de la fonction de fusion des notices individuelles.	138
15	Script Python : extrait de la fonction de création d'un corpus intermédiaire.	139
16	Script Python : extrait de la fonction de mise en conformité du corpus.	139
17	Script Python : extrait de la fonction d'enregistrement du corpus.	139
18	Application web : page d'accueil	144
19	Application web : liste des notices incorporées (axe 1).	145
20	Application web : science ouverte et documentation (axe 4).	146
21	Application web : index topographique.	147
22	Application web : mentions légales.	148
23	Workflow final.	150
24	Arborescence XML/TEI des notices.	152
25	Arborescence XML/TEI des notices.	153

Table des matières

Résumé	i
Remerciements	iii
Bibliographie	v
Développement applicatif et web	v
Encodage et gestion des données	v
Entités nommées et TAL	vii
Histoire et historiographie de la Ferme générale	viii
Humanités numériques et science ouverte	ix
Introduction	xiii
I Le <i>Dictionnaire de la Ferme générale</i> et l'ANR FermeGé, un <i>workflow</i> à diverses échelles	1
1 La Ferme générale comme objet d'étude	3
1.1 La Ferme générale comme objet historiographique	4
1.1.1 La Ferme générale : genèse, développement, remise en question	4
1.1.2 Un « modèle administratif » d'Ancien Régime ?	5
1.1.3 La Ferme générale : objet de contestations multiples	6
1.2 L'ANR et le projet FermeGé	8
1.2.1 L'ANR FermeGé : contexte, objectifs et fonctionnement	8
1.2.2 La Ferme générale : apports et enjeux historiographiques	9
1.3 La MESHS comme noeud d'un réseau infra-structurel	11
1.3.1 La MESHS : infrastructure d'accueil et d'appui à la recherche	11
1.3.2 Le réseau d'institutions mobilisées pour le projet	13
1.3.3 Le stage dans son contexte temporel et institutionnel	14
2 Encoder un dictionnaire historique en TEI	15
2.1 Panorama épistémologique de l'encodage des dictionnaires	16

2.1.1	Structuration de l'information dans un dictionnaire	16
2.1.2	Les dictionnaires et les métalangages d'encodage	17
2.1.3	Premières réflexions sur le <i>Dictionnaire de la Ferme générale</i>	18
2.2	Une approche linguistique dominante	19
2.2.1	De la linguistique au traitement automatique des langues	19
2.2.2	Les dictionnaires encodés en TEI : quelques projets d'encodage . .	20
2.3	Les <i>entry based documents</i>	23
2.3.1	L'hypothèse <i>Grobid-dictionnaries</i>	23
2.3.2	Des projets à la croisée des dictionnaires et bases de données	24
3	Premier encodage manuel prospectif	27
3.1	Rendre le sens explicite	28
3.1.1	La structuration conceptuelle des notices	28
3.1.2	Vers une esquisse d'un modèle logique	30
3.1.3	Conceptualiser les liens entre les notices	31
3.2	L'articulation des notices à l'échelle du dictionnaire	31
3.2.1	Un corpus de notices scientifiques ?	31
3.2.2	Le <i>Dictionnaire de la Ferme générale</i> : une oeuvre organique? . .	32
3.3	Les entités nommées : encodage, extraction et mise à disposition	33
3.3.1	Les entités nommées : l'apanage de la linguistique ?	33
3.3.2	Etat de l'art du traitement des entités nommées dans les sciences historiques	34
3.3.3	Les entités nommées et le projet FermeGé : quelles perspectives ? .	36
II	Du schéma à l'automatisation de l'encodage : mise en oeuvre d'une chaîne de traitement de la donnée	41
4	Structurer, schématiser et modéliser	43
4.1	L'O.D.D des notices individuelles	45
4.1.1	La schématisation de la structure des notices : le module <i>dictionaries</i> et ses limites	46
4.1.2	Les métadonnées et le teiHeader : de l'indexation à la pérennisation du projet	48
4.1.3	Sources et références scientifiques au sein des notices	51
4.2	L'O.D.D du corpus : le <i>Dictionnaire</i> comme une unité intellectuelle	53
4.2.1	De la nécessité (technique) d'un deuxième O.D.D : les méthodes « agiles » en action	53
4.2.2	D'une collection de notice à une oeuvre organique : un autre paradigme schématique	55

4.2.3	L'hypothèse de l' <i>O.D.D chaining</i> : complémentarité ou coexistence des schémas d'encodage ?	57
4.3	L'ontologie des entités nommées	58
4.3.1	Les enjeux scientifiques de l'encodage des entités nommées	58
4.3.2	Restreindre les valeurs des attributs par la grammaire O.D.D	60
4.3.3	Les limites de l'approche géographique	62
5	Le processus d'automatisation de l'encodage	65
5.1	Segmentation et encodage de la structure	66
5.1.1	Le choix des armes : traiter des données XML avec Python	66
5.1.2	Au début étaient les fonctions...	67
5.1.3	Des « arbres » à la notice : processus de fusion et mise en cohérence de l'arborescence XML/TEI	68
5.2	Encoder la bibliographie et les entités nommées	69
5.2.1	Un encodage de la structure par échelles	69
5.2.2	Différencier la bibliographie des sources	70
5.2.3	De nécessaires corrections : les limites de l'automatisation du processus	72
5.3	Une chaîne de transformation des documents	73
5.3.1	Appliquer le nouveau paradigme structurel : vers un processus d'encodage complémentaire	73
5.3.2	Mise en conformité des données du Dictionnaire	74
5.3.3	Vers l'application web : perspectives et limites du script d'encodage automatique	74
III	Le <i>Dictionnaire Numérique de la Ferme générale</i> : une esquisse de mise à disposition des données et ses enjeux	79
6	Mettre à disposition les données	81
6.1	L'application web : choix techniques et besoins scientifiques	83
6.1.1	D'un <i>Content Managing System</i> à une application web : des choix en lien avec l'avancement du projet	83
6.1.2	L'architecture : Flask comme <i>framework</i> de développement web	85
6.1.3	L'agencement graphique : Bootstrap comme <i>framework</i> graphique	87
6.2	État de l'art du Dictionnaire Numérique	89
6.2.1	Visualiser les notices et naviguer dans le Dictionnaire	89
6.2.2	Rechercher, requêter et explorer les notices ou index	90
6.2.3	Une application pour les lier tous : le site web comme confluence des axes du projet	91

6.3	Perspectives de développement de l'application web	92
6.3.1	De l'échantillon au Dictionnaire : l'intégration en continue des notices	92
6.3.2	La carte et le texte : anticiper les liens avec l' <i>Atlas de la Ferme générale</i>	93
6.3.3	Passer à l'échelle : vers l'hébergement et l'ouverture de l'application	97
7	Ouvrir les données	99
7.1	Le Dictionnaire et la « science ouverte »	101
7.1.1	La « science ouverte » : un nouveau paradigme pour la recherche en sciences humaines et sociales ?	101
7.1.2	La « science ouverte » : une pratique historique de partage des savoirs et techniques	102
7.1.3	La science ouverte et ses outils : le cas du <i>Dictionnaire Numérique de la Ferme générale</i>	104
7.2	La science ouverte et les enjeux éthiques ou juridiques	106
7.2.1	Une diversité de documents et illustrations : quels enjeux juridiques ?	106
7.2.2	Réutiliser les documents dans le cadre de l'enseignement supérieur et de la recherche	107
7.2.3	Outils et respect des enjeux éthiques et juridiques au sein du projet	108
7.3	Perspectives de développement	110
7.3.1	Développer une API : enjeux, processus et limites	110
7.3.2	Participer à la science ouverte : le cas des <i>data papers</i> et la « médiatisation » du projet	113
7.3.3	Réflexions sur le Dictionnaire, ses données et leur utilisation	115
Conclusion		119
Acronymes		123
Glossaire		125
Annexes		125
A. Dépôt Gitlab		129
.1	Arborescence	130
.2	Présentation des dossiers	131
B. Schématisation et documentation		135
.3	ODD des notices individuelles	135
.4	ODD du corpus de notices	135

.5	Ontologie et documentation	135
C. Scripts Python		137
.6	Script d'encodage des notices individuelles	137
.7	Script d'encodage du corpus TEI	138
D. Guide d'installation de l'application web « DicoNumFermeGé »		141
E. Application web « DicoNumFermeGé »		143
F. <i>Workflow</i> final		149
G. Arborescences XML/TEI		151
Table des figures		155
Table des matières		159