Humanités numériques

8 | 2023
Varia

Le *Dictionnaire numérique de la Ferme générale* : de la modélisation à la mise en ligne

The Dictionnaire numérique de la Ferme générale *: from Modelling to Online Publication*

Valentin De Craene et Victoria Le Fourner

Édition électronique
URL : https://journals.openedition.org/revuehn/3739
DOI : 10.4000/revuehn.3739
ISSN : 2736-2337

Éditeur
Humanistica

Référence électronique
Valentin De Craene et Victoria Le Fourner, « Le *Dictionnaire numérique de la Ferme générale* : de la modélisation à la mise en ligne », *Humanités numériques* [En ligne], 8 | 2023, mis en ligne le 01 décembre 2023, consulté le 02 mai 2024. URL : http://journals.openedition.org/revuehn/3739 ; DOI : https://doi.org/10.4000/revuehn.3739

humanités
numériques

8 | 2023
Varia

Le *Dictionnaire numérique de la Ferme générale* : de la modélisation à la mise en ligne

The Dictionnaire numérique de la Ferme générale : from Modelling to Online Publication

Valentin De Craene et Victoria Le Fourner

Résumés

Cette étude de cas présente le *Dictionnaire numérique de la Ferme générale (1664-1794)* créé lors du projet labellisé par l'Agence nationale de la recherche *FermeGé* en 2021, de sa modélisation conceptuelle jusqu'à son déploiement sur le Web. Si initialement le dictionnaire s'est imposé au collège d'historiens rédacteurs des notices comme un objet simple par sa structure formant une œuvre organique, la mise à disposition les données dans l'environnement du Web s'est révélée une étape essentielle dans la structuration de l'information. Ce point initial du dialogue entre les historiens et les ingénieurs a ouvert une réflexion sur la nécessité d'encoder en XML-TEI, d'une part, la structure des notices, d'autre part, les entités nommées afin d'ajouter une couche sémantique au texte brut. Dans cette même perspective, le traitement de la bibliographie et des sources mentionnées systématiquement en fin de notice a permis d'indexer ces éléments en vue d'un traitement ultérieur. Cette phase a amorcé un changement au sein du projet dans la manière d'appréhender le dictionnaire. Dès lors, il n'est plus uniquement une œuvre organique mais se transforme progressivement en un objet hybride adossé à une base de données relationnelle. Cet article entend donc analyser ce que les humanités numériques et l'ingénierie documentaire ont apporté à ce projet.

This case study presents the *Dictionnaire numérique de la Ferme générale (1664-1794)*, a digital dictionary of the General Farm created in the course of *FermeGé*, which is a project accredited by the Agence nationale de la recherche in 2021, from its conceptual modelling to its Web deployment.

While the dictionary initially appeared to the group of historians writing the entries to be a simple object with an organic structure, the need to make the data available in a Web environment proved to be an essential step in structuring the information. This initial point in the dialogue between historians and engineers led us to encode the structure of the records and the named entities in TEI XML to add a semantic layer to the raw text. In the same perspective, processing the bibliography and sources systematically mentioned at the end of a record enabled these elements to be indexed for later processing. This phase gradually changed the project's approach to the dictionary. From then on, the dictionary was no longer simply an organic work, but was gradually transformed into a hybrid object backed by a relational database. The aim of this article is to analyse the contribution of digital humanities and documentary engineering to the project.

Entrées d'index

MOTS-CLÉS : histoire, dictionnaire, encodage, TEI, entités nommées, Web

KEYWORDS: history, dictionary, encoding, TEI, named entities, Web

Introduction

1 Le dictionnaire est un objet couramment répandu dans les sciences historiques et *a priori* simple dans son organisation interne. Cependant, en termes de structuration de la donnée, il se révèle être un objet complexe et ambigu, dans la mesure où il est à la fois le support et un modèle de structuration de l'information. Le projet *FermeGé*, financé par l'Agence nationale de la recherche (ANR), s'est orienté vers la création d'un dictionnaire numérique afin de satisfaire une analyse historiographique « englobante » faisant appel à des chercheurs de différents horizons. Ce choix implique une réflexion sur la conceptualisation informatique d'un dictionnaire historique à dimension encyclopédique [1]. Effectivement, la perception du terme « dictionnaire historique » doit être replacée dans la galaxie des dictionnaires linguistiques numériques contemporains. À bien des égards, le *Dictionnaire numérique de la Ferme générale* prend la forme d'une encyclopédie faisant dialoguer différentes approches historiographiques et méthodes d'étude des sources. Or, la dimension linguistique des données produites au sein de ce « dictionnaire encyclopédique » est secondaire pour l'analyse historique développée par le collège d'historiens du projet. Ainsi, afin de saisir pleinement les spécificités du traitement des données au sein d'un dictionnaire historique encyclopédique numérique, il est nécessaire de comprendre l'articulation entre modélisation, encodage et publication de l'information produite. Dès lors, il nous faut positionner le projet face aux enjeux de modélisation conceptuelle et logique inhérents aux dictionnaires linguistiques dont il s'inspire largement dans sa structure informationnelle. En outre, le caractère nativement numérique du corpus de notices historiographiques produites pose la question des spécificités, si elles existent, d'un dictionnaire encyclopédique nativement numérique vis-à-vis d'une production papier. La notion de « nativement numérique » (*born digital*) recouvre l'ensemble des données, métadonnées et documents produits directement dans un contexte numérique, qu'il convient de différencier des sources numériques issues d'une étape préalable de numérisation et d'encodage (Paloque-Bergès 2015). À cet égard, il s'agit d'une première spécificité du projet ANR *FermeGé*, puisque, à la différence d'une chaîne de numérisation d'un dictionnaire préexistant, le travail principal d'ingénierie se déplace de la création et du contrôle qualité de l'océrisation vers la modélisation et l'encodage de l'œuvre. Dès lors, en quoi ce caractère encyclopédique du *Dictionnaire numérique de la Ferme générale* implique-t-il des enjeux de modélisation spécifiques ? De quelle manière la mise en œuvre d'une chaîne de traitement de la donnée entraîne-t-elle une progressive « lecture distante » du dictionnaire ?

2 En premier lieu, nous proposons d'expliciter plus en profondeur la genèse du projet *FermeGé* pour mettre en perspective le choix initial des historiens d'établir un dictionnaire numérique à dimension encyclopédique. Cette mise en contexte nous invite à présenter le modèle conceptuel sous-jacent du projet. Dans un second temps, nous étudions les choix de modélisation logique et d'encodage des données textuelles. Si les recommandations proposées dans le cadre du consortium de la Text Encoding Initiative (TEI) ont retenu notre attention, nous devons expliciter le positionnement de notre dictionnaire encyclopédique face à ces propositions de bali-

sage. Cette présentation de l'encodage nous permet de mettre en évidence l'importance de la chaîne de traitement de la donnée développée au cours du projet et de sa publication dans un environnement Web. Finalement, nous proposons de montrer la manière dont les questionnements de modélisation et de traitement des données ont permis de soulever de nouveaux enjeux en aval du projet. En effet, d'un livrable technique, le dictionnaire encyclopédique numérique devient lui-même un objet d'étude hybride, à la frontière entre une base de données et une œuvre organique.

Définir et conceptualiser l'objet : un « dictionnaire encyclopédique » de la Ferme générale

Genèse historiographique du projet et choix du dictionnaire numérique

3 Afin de saisir pleinement les raisons ayant poussé les membres du projet à choisir la forme d'un dictionnaire, il nous faut revenir un tant soit peu sur la genèse du projet et sa structuration. Le projet ANR *FermeGé : Administrer le privilège (1640-1794)* se veut la rencontre de diverses approches historiographiques complémentaires concernant la Ferme générale afin de proposer une analyse englobante et ainsi combler nombre de connaissances manquantes sur cette institution chargée de la perception des aides et autres taxes indirectes entre 1664 et 1794. Effectivement, il est indéniable que l'historiographie de la Ferme générale reste parcellaire [2]. Les premiers travaux, initiés notamment par l'ouvrage majeur de George Matthews en 1958, ont questionné la notion de rendement des taxes pour cette institution. En France, deux approches ont été privilégiées : une analyse socioculturelle des fermiers généraux (Durand 1971) et une approche plus judiciaire en lien avec les travaux sur la contrebande (Huvet-Martinet 1975). Jusqu'aux années 2000-2010, cette historiographie reste traversée par les questionnements sur la relation fisc et société dans la lignée de l'œuvre monumentale de Jean Nicolas (2002). Plus récemment, d'autres faisceaux d'analyse ont vu le jour, notamment autour des enjeux des circulations internationales, des techniques de gestion et d'ingénierie financière sous l'Ancien Régime et de l'apport de l'histoire du droit (histoire de la bureaucratie, du droit administratif et des juridictions [Boullu 2019]).

4 Ainsi, l'historien s'intéressant à la Ferme générale se trouve confronté à un paradoxe, disposant d'une multiplicité accrue et récente d'axes d'analyse, sans pour autant parvenir à restituer le rôle global de l'institution. C'est à partir de ce constat que se structurent la réflexion historiographique du projet *FermeGé* et sa méthode de travail. Les axes d'analyse conduisant à la rédaction du corpus de notices ont été établis à partir de notions et d'une terminologie propre à chaque approche historiographique. Ainsi, les historiens des lettres ont d'abord proposé une réflexion par terminologie autour des termes « exclusion », « inclusion », « manichéisme », amplifiée par les concepts propres aux historiens du droit. Face à cette grande diversité de thématiques naissantes et à l'ambition d'une histoire totale, le dictionnaire encyclopédique s'est imposé comme la

forme par excellence d'une production scientifique permettant de faire dialoguer des approches hétérogènes autour d'une épine dorsale : la notion de privilège.

5 À bien des égards, l'appellation de « dictionnaire » pour ce livrable scientifique et technique s'inscrit dans le paysage des dictionnaires historiques à l'usage des historiens, dont les enjeux sont sensiblement différents des dictionnaires linguistiques. Effectivement, le projet *FermeGé* s'insère dans la lignée des dictionnaires historiques tels que celui dirigé par Claude Gauvard pour le Moyen Âge (Gauvard, Libera et Zink 2002) ou de Lucien Bély pour l'Ancien Régime (Bély 1996), au sein desquels la définition est le fruit d'une analyse historique, fondée sur une bibliographie et des sources, développant ainsi une vision encyclopédique du sujet traité. Le choix de la production d'un dictionnaire encyclopédique s'est donc imposé, d'une part, comme un moyen de lier, comparer et prendre du recul sur un ensemble d'analyses historiographiques qu'une production papier n'aurait pas entièrement autorisé. D'autre part, les impératifs plus concrets de traitement et d'ouverture des publications et des données produites au sein des projets labellisés par l'ANR, à l'instar de *FermeGé* sur la période 2022-2025, ont poussé les porteurs du projet à opter pour une publication Web alignée sur les principes FAIR [3], grâce à l'appui technique de la Maison européenne des sciences de l'homme et de la société (MESHS). Or, le choix d'un livrable technique sous cette forme implique d'établir un modèle conceptuel solide, adapté au projet et à ses particularités.

Modélisation et conceptualisation d'un dictionnaire encyclopédique historique

6 Comme le soulignent Nancy Ide, Adam Kilgarriff et Laurent Romary, le préalable au développement concret d'un dictionnaire, tant numérique que papier, est la modélisation formelle de l'objet (Ide, Kilgarriff et Romary 2000). Du modèle conceptuel découlent donc un modèle logique, puis une implémentation physique au sein, dans notre cas, d'une application Web [4]. Ainsi, un dictionnaire encyclopédique peut être assimilé, au niveau conceptuel, à la structure mathématique d'un arbre, c'est-à-dire un graphe ordonné, dirigé et, en théorie, acyclique. En d'autres termes, les dictionnaires, dont nous nous inspirons pour établir notre modèle, sont des arbres comprenant un élément racine englobant (correspondant au niveau le plus élevé du corpus de notices), à partir duquel descendent des branches (représentant les notices et les blocs d'informations caractérisant la forme ou le sens). Ces branches contiennent à leur tour des segments transmettant une information spécifique, qu'elle soit linguistique ou, plus précisément dans notre cas, sémantique et bibliographique.

7 C'est au niveau des segments d'information que les dictionnaires se distinguent du modèle mathématique des arbres, en raison des éléments de « mésostructure » (Lemnitzer, Romary et Witt 2013) ou de références croisées. Concrètement, au sein d'une entrée de dictionnaire ou d'encyclopédie peuvent être référencées d'autres entrées, rompant ainsi la dimension acyclique du modèle. Dans ces conditions, il nous faut opter pour un modèle d'encodage prenant en compte ce mécanisme de références croisées en leur attribuant des identifiants uniques. En outre, en suivant ce modèle

conceptuel de l'arbre, les informations sont propagées au sein du dictionnaire suivant le principe global d'héritage. Une information contenue à un niveau supérieur se propage et s'applique aux niveaux inférieurs.

8 Dans le cas du *Dictionnaire numérique de la Ferme générale*, soulignons le fait que les informations au niveau des titres des entrées sont « remplaçantes » car elles se soustraient aux informations précédentes, alors que les définitions sont cumulatives, puisqu'elles s'ajoutent aux informations précédemment renseignées au niveau supérieur (en l'occurrence le titre). Ainsi, nous proposons de synthétiser le modèle conceptuel que nous avons adopté sous forme d'un arbre schématique (figure 1).

Figure 1. Modèle conceptuel simplifié du *Dictionnaire numérique de la Ferme générale*

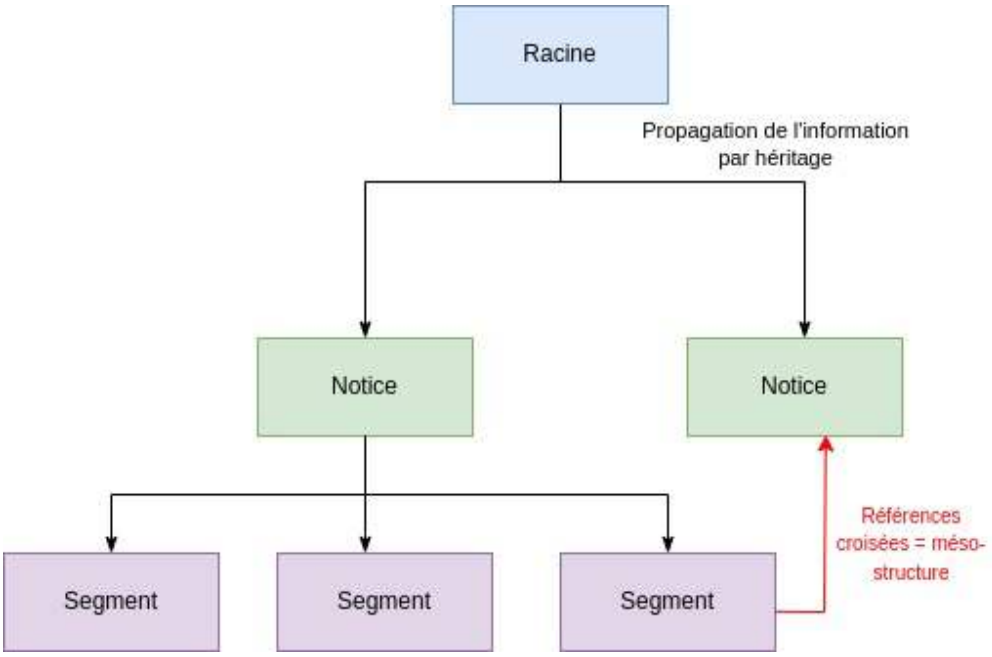

Figure produite par les auteurs

9 Cette représentation conceptuelle du dictionnaire numérique soulève une question que nous devons examiner et approfondir par la suite : s'agit-il d'une œuvre organique dont la racine englobe toutes les entrées ou d'un ensemble de notices présentant une racine distincte pour chaque entrée ? Afin de clarifier le traitement de notre modèle conceptuel, replaçons le projet *FermeGé* dans le contexte de l'élaboration et du traitement des dictionnaires numériques contemporains, principalement linguistiques.

Quel positionnement pour un dictionnaire encyclopédique historique face aux dictionnaires linguistiques ?

10 La modélisation et l'encodage des données au sein des dictionnaires ont intéressé historiquement au premier chef les linguistes. Comme le rappelle Jean Pruvost (2010), deux dynamiques fondatrices sont au cœur de l'étude et du traitement des dictionnaires : d'une part, la « métalexicographie », à la suite des travaux de Bernard Quemada, qui allie la lexicographie (l'étude et le recensement des mots d'une langue) et la dictionnairique (la « fabrique » des dictionnaires à destination d'un public), s'intéressant donc au sens des mots, et, d'autre part, la « lexiculture », initiée par Robert Galisson, qui étudie les représentations culturelles portées par les mots. Dans cette optique, il n'est pas surprenant de noter que les modèles d'encodage des dictionnaires se sont principalement développés au sein des communautés de chercheurs en linguistique. À cet égard, la TEI et son module « *dictionaries* » cristallisent de nombreux enjeux de structuration de la donnée et illustrent cette situation. Sur ce point, l'article publié par Ide et Véronis en 1995, présentant l'élaboration du module dans la revue *Computers and the Humanities*, est limpide quant aux champs disciplinaires visés, à savoir par ordre d'importance : l'édition numérique, la lexicographie, la linguistique, la philologie et enfin l'histoire de l'imprimé (Ide et Véronis 1995, 14-15). Dans cette perspective, le module « *dictionaries* » propose un balisage adapté à une analyse étymologique ou grammaticale. La TEI est donc conforme à la vision « sémasiologique » de la structure lexicale, puisqu'elle s'intéresse principalement au sens et organise les entrées de manière séquentielles et hiérarchiques.

11 De même, se pose la question de son succès et de ses possibles limites pour l'encodage des dictionnaires, d'une part, et de notre encyclopédie historique, d'autre part. Si Mangeot et Enguehard (2013, 8) soulignent que le succès de l'encodage des dictionnaires en TEI n'est que partiel, force est de constater que des initiatives plus récentes de mutualisation des moyens s'appuyant sur des corpus TEI ont vu le jour. Citons tout spécifiquement l'initiative de spécification de la TEI destinée aux ressources lexicographiques intitulée TEI Lex-0. Ensemble de spécifications techniques fondées sur les *Recommandations* de la TEI et l'expérience de la communauté des utilisateurs, la TEI Lex-0 a pour objectif d'améliorer l'interopérabilité des ressources lexicographiques encodées et de produire des dictionnaires plus aisément requêtables en rendant obligatoires certains éléments ou attributs (Tasovac *et al.* 2018). De même, l'infrastructure européenne Elexis de mutualisation des outils lexicographiques, fondée sur un modèle et un vocabulaire commun, a pour objectif de fournir aux chercheurs en traitement automatique du langage naturel (TALN, *natural language processing* [NLP], en anglais) et en intelligence artificielle (IA) des données lexicographiques de haute qualité (Tiberius *et al.* 2021).

12 Ce constat dressé, il nous faut à présent étudier les recommandations d'encodage des dictionnaires intéressant le projet *FermeGé* et contenues dans le chapitre « *dictionaries* » des *Recommandations pour l'encodage et l'échange des textes électroniques* de la Text Encoding Initiative (2023). Rappelons que ce chapitre, initialement intitulé « *Print dictionaries* » (Lemnitzer, Romary et Witt 2013, 12), se veut un compromis entre un format hautement structuré et un outil permissif de balisage de certains diction-

naires anciens peu structurés. Cette situation explique la présence d'un arsenal de différents éléments pour l'encodage des entrées de dictionnaires, parmi lesquels nous devons faire des choix pour correspondre, d'une part, à notre modèle conceptuel et, d'autre part, aux enjeux du projet en matière de traitement, publication et ouverture des données.

De la modélisation à l'encodage : mise en place d'une chaîne de traitement de la donnée

13 En somme, le choix d'un dictionnaire encyclopédique a été motivé par la nécessité de faire dialoguer au sein d'un même corpus un ensemble de notices relevant d'approches différentes et complémentaires. Si d'un point de vue conceptuel, le modèle d'un arbre a été retenu, les choix d'encodage en TEI, s'insérant dans une chaîne de traitement plus vaste, doivent à présent être explicités.

Enjeux et exemples d'encodage du dictionnaire encyclopédique en TEI

14 L'encodage des notices du *Dictionnaire numérique de la Ferme générale* repose donc principalement sur le module « *dictionaries* » de la TEI, tirant profit de sa flexibilité. Effectivement, comme nous ne traitons pas des informations lexicales et étymologiques au sein des notices scientifiques, nous n'avons eu recours qu'à un jeu d'éléments relativement restreint que nous pouvons classer en quatre catégories.

- Les éléments de structure, qui portent des informations principalement cumulatives à l'échelle des notices et qui mettent en exergue les niveaux hiérarchiques d'informations, à savoir : <entry [5]>, contenant, d'une part, <form [6]> qui indique la forme prise par le lemme et, d'autre part, <sense [7]> contenant lui-même <def [8]> développant « en prose » l'analyse scientifique et historiographique. Le choix de l'élément <entry> a été déterminé par la récurrence structurelle de l'ensemble des notices, au détriment d'<entryFree [9]> proposant une plus grande flexibilité qui ne nous semblait pas nécessaire dans le cadre de la génération d'un modèle logique strict (Budin, Majewski et Mörth 2012).
- Les éléments relevant de la « mésostructure » du dictionnaire encyclopédique, constitués des renvois entre les notices, que nous avons encodés avec les éléments <ref [10]>. L'ajout d'un attribut @target permet de pointer vers une autre notice unique et ainsi gérer, comme nous le montrons plus bas, les références croisées grâce à une transformation en Extensible Stylesheet Language Transformation (XSLT).
- Les éléments de bibliographie, séparés en trois catégories : les sources primaires mobilisées pour la rédaction de la notice, les sources imprimées et les références bibliographiques. Ces références sont contenues dans un élément englobant <listBibl [11]> puis dans des éléments enfants <bibl [12]> que la valeur des attributs @type permet de différencier suivant la méthodologie de classification des sources

abordées. Cela évite une imbrication des éléments <bibl> telle que cela avait été initialement envisagé, pour correspondre au mieux au modèle conceptuel.

- Les métadonnées qui sont recensées dans le <teiHeader [13]> en accord avec les recommandations de la TEI. Les références bibliographiques auxquelles se rattachent les notices ont été dans ce cas indexées au sein des éléments <profileDesc [14]> et <textClass [15]> en respectant le vocabulaire contrôlé du Répertoire d'autorité matière encyclopédique et alphabétique unifié (RAMEAU [16]).

15 Le cas de l'encodage des entités nommées mérite un tant soit peu notre attention. En encodant ces entités, nous ouvrons la perspective de dépassement de la condition de dictionnaire pour tendre vers la création d'une base de données des entités nommées relatives à la Ferme générale. À cet égard, la possibilité d'indexer ce corpus de lieux et d'institutions puis de formuler des requêtes invite à explorer le dictionnaire non plus simplement par les clés (les titres) mais directement par les valeurs (le contenu des définitions). Dans cette perspective, l'équipe de coordination de l'axe de recherche a donc développé une typologie d'entités qui a fait l'objet d'un *mapping* en TEI. Rappelons brièvement que la notion même d'entité nommée est relativement permissive et donne lieu à diverses réflexions et débats, notamment au sein du consortium de la TEI, sur la place par exemple des entités « abstraites », faisant référence à des personnalités morales, des personnages pseudo-historiques ou des allégories et divinités clairement nommées (Jurafsky et Martin 2022, 25). Or, l'une des ambitions majeures du projet *FermeGé* est de comprendre et d'analyser l'emprise spatiale et multiscalaire de cette institution complexe qu'est la Ferme générale vis-à-vis des territoires et de la société d'Ancien Régime. Cet objectif a motivé l'établissement d'une typologie des entités nommées de lieux et d'organisations que le dictionnaire encyclopédique se doit d'encoder. Dans la perspective de s'inscrire dans une démarche d'ouverture des données, les entités nommées indexées automatiquement sont alignées sur un ensemble de référentiels pérennes tels que les identifiants DicoTopo [17]. Si ces derniers sont particulièrement bien adaptés au référencement des toponymes de la France d'Ancien Régime, ils ne couvrent à ce jour qu'un tiers du royaume et ne concernent que très rarement des institutions relevant de plusieurs territoires administratifs tels que les généralités ou les diocèses. Dès lors, cette phase d'alignement des données est complétée par l'ajout des référentiels GeoNames [18] ou d'identifiants Wikidata [19] lorsque DicoTopo fait défaut. En conséquence, cet alignement des entités nommées permet de s'inscrire dans une démarche de décloisonnement des données produites et ainsi de dépasser l'indexation d'un simple dictionnaire papier.

16 Le schéma ci-dessous présente donc une adaptation du modèle conceptuel aux choix d'encodage en XML-TEI du *Dictionnaire numérique de la Ferme générale*. Les relations représentées par les flèches sont hiérarchiques (un élément « contient » d'autres éléments, voir figure 2).

Figure 2. Modèle logique (arborescence XML-TEI) du *Dictionnaire numérique de la Ferme générale*

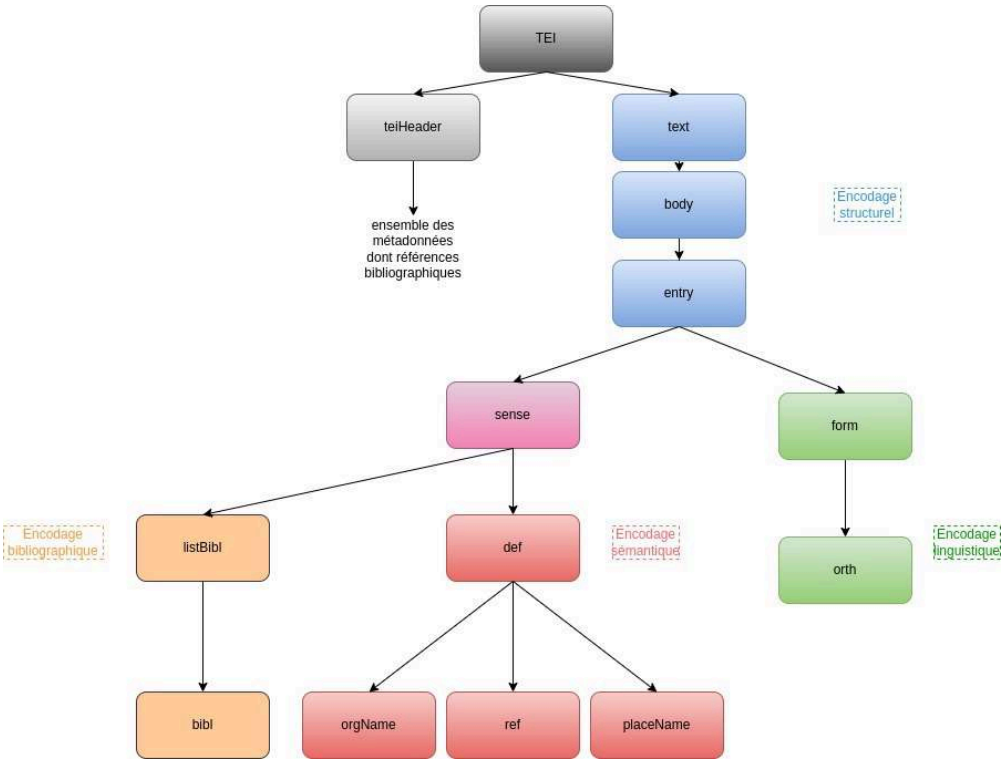

Figure produite par les auteurs

Une chaîne d'encodage et de traitement de la donnée pour le projet

17 Afin de mettre en œuvre ce balisage en TEI à l'échelle du dictionnaire, il est nécessaire de développer une chaîne d'encodage semi-automatisée. En effet, le corpus initial rédigé par les historiens s'établit, début 2023, à 275 notices, ce qui représente environ 500 pages de texte lorsqu'elles sont compilées. Dans le cadre du projet *FermeGé*, un encodage manuel n'est donc pas envisageable, pour des raisons techniques et infrastructurelles évidentes. Par conséquent, nous avons mis en place une chaîne d'encodage permettant de traiter l'ensemble des notices avec une granularité jugée satisfaisante pour le projet. Cette chaîne a été développée en Python en utilisant les bibliothèques ElementTree et Beautiful Soup 4 pour la manipulation et la structuration des données XML [20]. En entrée, l'algorithme prend les notices au format texte brut, puis insère les balises correspondant à l'encodage structurel et sémantique, grâce à un ensemble d'expressions régulières. D'un point de vue technique, le script « capture » les chaînes de caractères recherchées, insère les balises ouvrantes et fermantes, puis « parse » un arbre XML créé pour chaque élément. L'ensemble des fragments d'arbre sont ensuite réinsérés hiérarchiquement et « parsés » afin de produire une notice encodée valide et conforme aux besoins de notre encodage, schématisée au sein d'un document ODD (« *one document does it all* ») propre au projet [21].

18 Éminemment imprégné des principes du *literate programming* (Knuth 1984), l'ODD présente ce double avantage de réunir en un seul et même document le schéma technique d'encodage, destiné à être transformé en Relax NG (Regular Language for XML Next Generation), et la documentation « en prose ». Afin de repérer les éventuels arbres XML-TEI erronés ou mal

insérés au cours du traitement, des règles supplémentaires et restrictives dans le langage Schematron ont été ajoutées. L'ODD est donc la clé de voûte de cet édifice de traitement de la donnée puisqu'elle permet de valider le corpus qui est ensuite visualisé au sein d'une application Web grâce à un ensemble de règles de transformation XSLT. L'intérêt de XSLT pour notre projet est double : il permet, d'une part, de fournir, par le biais d'une feuille de transformation, une sortie encodée en HTML des notices et du corpus dans son ensemble, qui peut ensuite être visualisée au sein de l'application Web, et, d'autre part, de produire des sorties dans d'autres formats, utilisées au sein du projet pour d'autres objectifs, tels que la possibilité de générer des fichiers LaTeX [22], destinés à être transformés au format PDF. Nous proposons de synthétiser l'articulation entre notre chaîne de traitement de la donnée et la mise en valeur postérieure au sein de l'application Web du projet par le biais de la figure 3.

Figure 3. Schéma de la chaîne de traitement et publication du *Dictionnaire numérique de la Ferme générale*

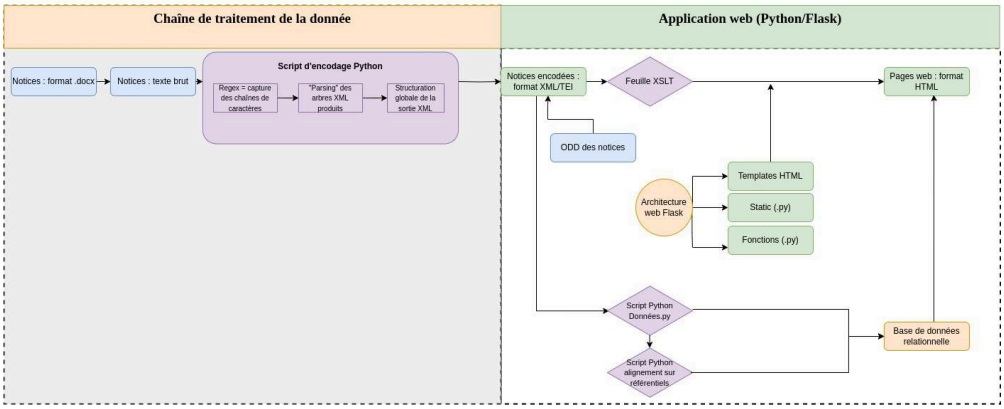

Figure produite par les auteurs

19 Soulignons l'utilisation du script intitulé « Données.py » qui permet, par le biais d'expressions XPath, de générer une base de données relationnelle recensant les notices, les auteurs, les éléments bibliographiques, les sources primaires et imprimées, ainsi que les entités nommées associées à leurs identifiants pérennes. Ce script parcourt l'ensemble de l'arborescence des notices pour capturer les chaînes de caractères demandées, puis crée une « session » permettant de mettre à jour la base de données. C'est à partir de cette base que peuvent être proposés des moteurs de recherche dits « à facettes », qui croisent les requêtes. Cette phase postérieure à la chaîne de traitement de la donnée tire profit de l'interopérabilité de la TEI pour mettre en œuvre un ensemble de fonctionnalités dépassant la simple publication Web au format HTML. Dans cette même perspective, cette multiplication des points d'entrée au sein du dictionnaire encyclopédique et numérique implique une forme de basculement de la conception même de ce livrable technique.

Du corpus au dictionnaire : vers un objet hybride et un nouveau paradigme au sein du projet ?

20 Effectivement, la possibilité de naviguer au sein du corpus se voit décuplée, que ce soit par les hyperliens entre les notices, les recherches par thématiques historiographiques, par éléments bibliographiques communs ou par toponymes et organisations. En conséquence, nous percevons ici le véritable apport du traitement numérique du dictionnaire encyclopédique, à savoir la capacité à dépasser la structuration initiale de l'information. De fait, au sein de notre modèle logique, l'accès à l'information se fait principalement par le couple entrée – définition. Or, par ces différents moyens de « requêter » le corpus, nous renversons la perspective du dictionnaire, ce qui entraîne la possibilité de mettre en place une forme de « lecture distante » appliquée à la Ferme générale.

21 Au stade actuel du projet, une forme d'analyse distante et englobante du *Dictionnaire numérique de la Ferme générale* se structure autour des différents outils d'exploration que nous avons développés. Ce changement d'échelle, en passant du niveau de la notice à celui du corpus numérique, pose la question de la nature même du dictionnaire. En effet, la constitution d'une véritable base de données relationnelle des entités nommées, ainsi que l'indexation multiscalaire des notices, invitent à s'interroger sur le caractère organique du dictionnaire. Cette phase de développement du projet indique un changement de paradigme vis-à-vis du *Dictionnaire numérique de la Ferme générale*, puisque, par la lecture distante, il n'est plus simplement le support de l'information scientifique mais devient un objet de réflexion à part entière. La lecture distante (*distant reading*) en tant qu'approche des données textuelles a pour finalité de rechercher et d'identifier la présence de motifs (*patterns*) récurrents au sein d'un corpus (Puren 2020). À l'instar de l'étude fondatrice de Moretti (2009) sur les titres des romans britanniques entre 1740 et 1850, cette méthode entend remettre dans leur contexte les textes canoniques, cette approche semble transposable au contenu des notices du dictionnaire. Ainsi, la mise en réseau par l'indexation et la visualisation en graphe nous permet d'identifier la présence de thématiques récurrentes et communes à certains groupes de notices. Les notions de contestation, de privilège, de vexation, de compromis et de fiscalité sont parmi les plus courantes, même si elles restent encore difficilement quantifiables. Dès lors, cette approche englobante et distante du contenu des notices renforce à nouveau le dialogue entre ingénieurs et historiens puisqu'il importe de circonscrire historiographiquement ces thématiques que l'approche distante révèle.

22 Dans cette perspective, ces thématiques sont encodées et indexées afin de permettre une recherche avancée au sein de ce corpus. Cette phase d'encodage des notions et thématiques au cœur des notices du dictionnaire s'avère être une phase supplémentaire de transformation du dictionnaire initial en un objet nouveau et hybride. Cette déconstruction de l'objet dictionnaire par l'indexation automatisée des notions historiographiques s'insère dans un processus plus large que l'historien Lincoln Mullen (2018) a pu qualifier de « récit (historique) tressé ou brodé » (*braided narratives*). Cette forme nouvelle d'écriture scientifique, que Mullen ap-

pelle à généraliser, tend à associer l'analyse scientifique et le traitement de la donnée ou la méthodologie de structuration adoptée dans un seul et même type de récit. L'analyse scientifique n'est donc nullement décorrélée du traitement des données. Ainsi, le *Dictionnaire numérique de la Ferme générale* s'affirme comme le résultat d'un dialogue fertile entre historiens et ingénieurs, donnant lieu à l'émergence d'un objet hybride qui dépasse la simple condition initiale du dictionnaire traditionnel. Cette lecture distante a pour objectif d'être réintégrée au sein du projet par le biais de visualisations de données interactives, comme la visualisation et la navigation au sein du dictionnaire à partir d'un graphe de connaissances (figure 4).

Figure 4. Visualisation par graphe des citations internotices (extrait, lettre A)

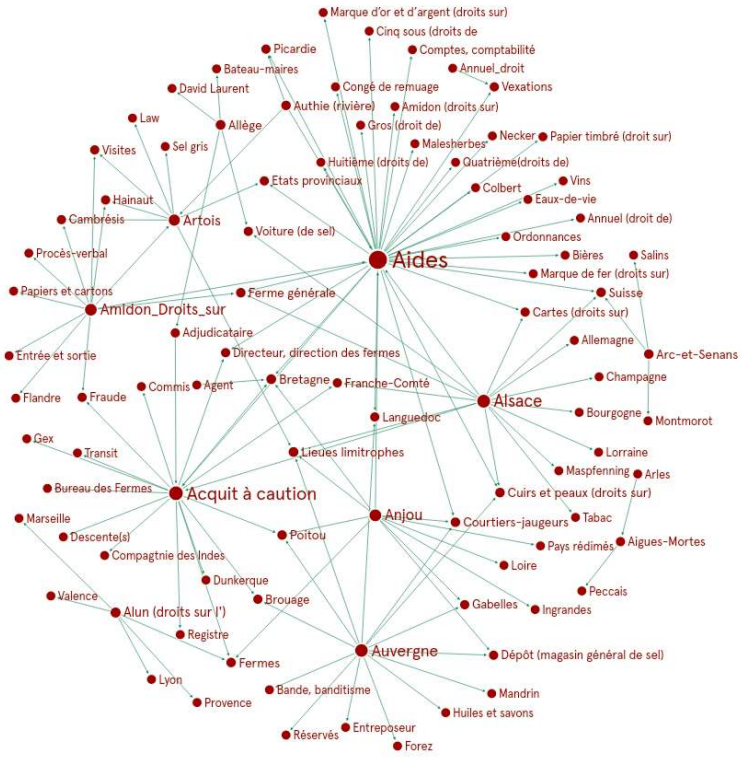

Figure produite par les auteurs

23 Pour une question de lisibilité et de représentativité, nous ne fournissons ici qu'un extrait réduit du graphe concernant la lettre A. En ayant recours à un service externe [23] dans lequel nous importons la matrice du graphe sous un format comma-separated values (CSV), nous sommes en mesure de visualiser les références croisées entre les notices.

Conclusion

24 D'un modèle conceptuel, largement fondé sur celui d'un dictionnaire, au développement d'une application Web permettant d'explorer, de visualiser et « requêter » notre *Dictionnaire numérique de la Ferme générale*, la question de l'encodage des données reste centrale. Nous avons eu recours principalement au module « *dictionaries* » des *Recommandations* de la TEI, permettant un encodage minimal de la structure et des données linguistiques, nous focalisant ainsi sur l'encodage sémantique des entités nommées et le traitement des références croisées. En intégrant ce corpus dans une application Web, nous observons un changement dans la conception du projet. D'un livrable technique, le dictionnaire encyclopédique numérique devient un objet de réflexion, proposant une lecture distante ainsi qu'une visualisation d'un corpus et des questionnements historiographiques associés à la Ferme générale. En somme, cette approche itérative du traitement des données permet de renforcer le dialogue entre historiens et ingénieurs, que ce soit en amont du projet, dans la conception du modèle d'encodage, puisqu'il implique une réflexion sur la structure des notices, ou en aval du traitement, par les fonctionnalités numériques offertes de « requêtage » et de visualisation. Ainsi, nous ne pouvons qu'encourager le développement du dialogue entre historiens et ingénieurs en amont de la rédaction scientifique afin de s'inscrire dans une perspective facilitant le traitement et la pérennisation des données de la recherche. Concluons en relayant les propos de l'historien Olivier Poncet à propos de la thèse de Thierry Claes, intitulée *Dictionnaire biographique des financiers en France au XVIII^e^ siècle*, qui déclarait à cet égard que « tout chercheur devrait voir son travail être détourné à des fins qu'il n'a pas imaginées » (Poncet 2009). Nous espérons, par cette déconstruction et l'émergence d'une nouvelle forme d'objet à partir du *Dictionnaire numérique de la Ferme générale*, permettre ce détournement à des fins heureuses du travail collectif des historiens.

Bibliographie

Bély, Lucien, éd. 1996. *Dictionnaire de l'Ancien Régime*. Paris : Presses universitaires de France.

Boullu, Thomas. 2019. « La transaction en matière d'impositions indirectes (1661-1791) : contribution à l'émergence d'un droit de l'administration monarchique ». Thèse de doctorat en histoire du droit et des institutions, université de Strasbourg.

Budin, Gerhard, Stefan Majewski et Karlheinz Mörth. 2012. « Creating Lexical Resources in TEI P5 ». *Journal of the Text Encoding Initiative* 3. https://doi.org/10.4000/jtei.522.

Durand, Yves. 1971. *Les Fermiers généraux au XVIII^e^ siècle*. Paris : Presses universitaires de France.

Gauvard, Claude, Alain de Libera et Michel Zink, éd. 2002. *Dictionnaire du Moyen Âge*. Paris : Presses universitaires de France.

Huvet-Martinet, Micheline. 1975. « Gabelous et faux-sauniers en France à la fin de l'Ancien Régime. Essai statistique et sociologique sur le faux-saunage dans le ressort de la Commission de Saumur (1764-1789) ». Thèse de 3^e^ cycle en histoire, université de Rennes 2.

Ide, Nancy, Adam Kilgarriff et Laurent Romary. 2000. « A Formal Model of Dictionary Structure and Content ». Dans *Proceedings of Euralex 2000*. Leiden : European Association for Lexicography. https://hal.science/hal-00164625.

Ide, Nancy et Jean Véronis. 1995. « Encoding Dictionaries ». *Computers and the Humanities* 29 (mars) : 167-179. https://doi.org/10.1007/BF01830710.

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

Knuth, Donald Ervin. 1984. « Literate Programming ». *The Computer Journal* 27 (2) : 97-111. https://doi.org/10.1093/comjnl/27.2.97.

Legay, Marie-Laure. 2022. « Le dictionnaire de la Ferme générale. Workshop du 6 janvier 2022 ». *Le projet FermeGé : administrer le privilège (1640-1794)* (blog). https://dicofg.hypotheses.org/3556/.

Lemnitzer, Lothar, Laurent Romary et Andreas Witt. 2009. « Representing Human and Machine Dictionaries in Markup Languages ». Dans *Dictionaries. An International Encyclopedia of Lexicography. Supplementary Volume : Recent Developments with Special Focus on Computational Lexicography*, édité par Rufus Gouws, Ulrich Heid, Wolfgang Schweickard et Herbert Ersnt Wiegand, 1195-1208. Berlin et Boston : de Gruyter Mouton. https://inria.hal.science/inria-00441215.

Mangeot, Mathieu et Chantal Enguehard. 2013. « Des dictionnaires éditoriaux aux représentations XML standardisées ». Dans *Ressources lexicales. Contenu, construction, utilisation, évaluation*, édité par Núria Gala et Michael Zock. Amsterdam : John Benjamins. https://doi.org/10.1075/lis.30.08man.

Matthews, George T. 1958. *The Royal General Farms in Eighteenth-Century France*. New York : Columbia University Press.

Moretti, Franco. 2009. « Style, Inc. Reflections on Seven Thousand Titles (British Novels, 1740-1850) ». *Critical Inquiry* 36 (1) : 134-158. https://doi.org/10.1086/606125.

Mullen, Lincoln. 2018. « A Braided Narrative for Digital History » Dans *Debates in the Digital Humanities 2019*, édité par Matthew K. Gold et Lauren F. Klein, 606-617. Minneapolis : University of Minnesota Press. https://hcommons.org/deposits/item/hc:18095/.

Nicolas, Jean. 2002. *La Rébellion française. Mouvements populaires et conscience sociale 1661-1789*. Paris : Le Seuil.

Paloque-Bergès, Camille. 2015. « Les sources nativement numériques pour les sciences humaines et sociales ». *Histoire@Politique* 30 (3) : 221-244. https://doi.org/10.3917/hp.030.0221.

Poncet, Olivier. 2009. « Thierry Claes. *Dictionnaire biographique des financiers en France au XVIII^e siècle* ». *Bibliothèque de l'École des chartes* 167 (2) : 585-586. https://www.persee.fr/doc/bec_0373-6237_2009_num_167_2_463981_t15_0585_0000_1.

Pruvost, Jean. 2010. « La traque lexicographique et dictionnairique : du loup au chat en passant par le vin, le mariage et le citoyen ». *Éla. Études de linguistique appliquée* 157 (1) : 103-110. https://doi.org/10.3917/ela.157.0103.

Puren, Marie. 2020. « La lecture distante : introduction et exemples d'application ». Matériaux pédagogiques, Master 1 EMAS, université de Versailles – Saint-Quentin-en-Yvelines. https://hal.archives-ouvertes.fr/hal-03152747.

Rey, Alain. s. d. « Encyclopédie ». *Encyclopædia Universalis*. https://www.universalis.fr/encyclopedie/encyclopedie/.

Tasovac, Toma, Laurent Romary, Piotr Banski, Jack Bowers, Jesse de Does, Katrien Depuydt, Tomaž Erjavec, Alexander Geyken, Axel Herold, Vera Hildenbrandt, Mohamed Khemakhem, Boris Lehečka, Snežana Petrović, Ana Salgado et Andreas Witt. 2018. « TEI Lex-0 : A baseline encoding for lexicographic data. Version 0.9.2 ». *DARIAH Working Group on Lexical Resources*. https://dariah-eric.github.io/lexicalresources/pages/TEILex0/TEILex0.html.

Text Encoding Initiative. 2023. « Dictionaries ». *Recommandations pour l'encodage et l'échange des textes électroniques*. Version 4.6.0. https://tei-c.org/release/doc/tei-p5-doc/fr/html/DI.html.

Tiberius, Carole, Simon Krek, Katrien Depuydt, Polona Gantar, Jelena Kallas, Iztok Kosem et Michael Rundell. 2021. « Towards the ELEXIS Data Model : Defining a Common Vocabulary for Lexicographic Resources ». Dans *Proceedings of eLex Conference 2021*. https://zenodo.org/records/7118651.

Notes

1 Nous comprenons le terme d'encyclopédie comme une œuvre de référence entendant synthétiser l'ensemble des connaissances disponibles à un instant t sur un sujet donné, s'intéressant plus aux définitions et contenus scientifiques qu'aux mots et au contexte linguistique. Comme le soulignait Alain Rey, « l'encyclopédie est ouverte sur le monde » : https://www.universalis.fr/encyclopedie/encyclopedie/.

2 La genèse du projet, associée à un bilan historiographique plus dense, est développée en détail par Marie-Laure Legay, directrice scientifique du projet, dans le compte rendu du séminaire qui s'est tenu à l'université de Nanterre le 6 janvier 2022 (Legay 2022).

3 Principes fondamentaux d'ouverture et d'interopérabilité des données (FAIR pour « facile à trouver, accessible, interopérable et réutilisable »).

4 https://fermege.meshs.fr.

5 https://tei-c.org/release/doc/tei-p5-doc/en/html/ref-entry.html.

6 https://tei-c.org/release/doc/tei-p5-doc/en/html/ref-form.html.

7 https://tei-c.org/release/doc/tei-p5-doc/en/html/ref-sense.html.

8 https://tei-c.org/release/doc/tei-p5-doc/en/html/ref-def.html.

9 https://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-entryFree.html.

10 https://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-ref.html.

11 https://tei-c.org/release/doc/tei-p5-doc/en/html/ref-listBibl.html.

12 https://tei-c.org/release/doc/tei-p5-doc/fr/html/ref-bibl.html.

13 https://tei-c.org/release/doc/tei-p5-doc/fr/html/ref-teiHeader.html.

14 https://tei-c.org/release/doc/tei-p5-doc/fr/html/ref-profileDesc.html.

15 https://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-textClass.html.

16 Formulée en langage naturel et décrivant le contenu des notices en termes d'approche historique et historiographique (« histoire financière », « histoire des institutions », « histoire des contestations »), cette typologie est transcrite au sein des métadonnées dans le <teiHeader> dans le langage RAMEAU maintenu et mis à jour par la Bibliothèque nationale de France. Bien que RAMEAU fasse l'objet d'une réforme dans sa structure, il nous semble qu'il s'agit d'un moyen d'assurer la pérennité du référencement du dictionnaire sur le moyen et long terme.

17 Une introduction au projet *DicoTopo* se trouve à l'adresse suivante : https://dicotopo.cths.fr/about/.

18 https://www.geonames.org.

19 https://www.wikidata.org/wiki/Wikidata:Main_Page.

20 L'ensemble de l'application Web (/app) et de la chaîne de traitement (/utils) se trouve à l'adresse suivante : https://gitlab.huma-num.fr/vdecraene/DicoNumFermeGe/.

21 https://gitlab.huma-num.fr/vdecraene/DicoNumFermeGe/-/blob/main/ODD_schemas/ODD_v2.xml.

22 Dans ce cas, nous utilisons une sortie au format texte avec une extension .tex. Voir https://gitlab.huma-num.fr/vdecraene/DicoNumFermeGe/-/blob/main/utils/XML_to_TeX/XMLtoLaTeX.xsl/.

23 Initialement, nous avons eu recours à l'outil Graph Commons (https://graphcommons.com) sur le modèle de l'*Ontologie du christianisme médiéval en images* développé par l'Institut national d'histoire de l'art (https://omci.inha.fr/s/ocmi/page/omcigraph/), mais nous basculons progressivement vers Gephi Lite (https://gephi.org/gephi-lite/).

Auteurs

Valentin De Craene

UAR 3185 MESHS, projet ANR *FermeGé*, Lille, France
Agrégé d'histoire et diplômé du master Technologies numériques appliquées à l'histoire de l'École nationale des chartes, Valentin De Craene est ingénieur d'études contractuel au sein du projet *FermeGé*, financé par l'Agence nationale de la recherche et géré par la Maison européenne des sciences de l'homme et de la société (MESHS) et l'université de Lille.
ORCID 0000-0002-6019-2572
valentin.de-craene@univ-lille.fr

Victoria Le Fourner

UAR 3185 MESHS, CNRS, Lille, France
Ingénieure d'études au CNRS, Victoria Le Fourner travaille à la Maison européenne des sciences de l'homme et de la société (MESHS) à Lille, après un parcours en humanités numériques.
ORCID 0000-0002-8053-0356
victoria.le-fourner@univ-lille.fr

Droits d'auteur