Actes des journées humanités numériques et Web sémantique

édités par Nicolas Lasolle, Olivier Bruneau et Jean Lieber

Nancy
LORIA, 21 et 22 juin 2021
Archives Henri-Poincaré, 21 juin 2022

Sommaire

Comment référencer ce volume

Nicolas Lasolle, Olivier Bruneau et Jean Lieber, éd. (juin 2022). *Actes des journées humanités numériques et Web sémantique* (Nancy, France). DOI : 10.5281/zenodo.7014341. URL : https://humanumwebsem.sciencesconf.org/

Comment référencer un chapitre

Format général

Prénom Nom (juin 2022). « Titre de l'article ». *Actes des journées humanités numériques et Web sémantique* (Nancy, France). Sous la dir. de Nicolas Lasolle, Olivier Bruneau et Jean Lieber, p. n-m. DOI : 10.5281/zenodo.7014341

Rendu pour le premier article

Francesco Beretta (juin 2022). « Interopérabilité des données de la recherche et ontologies fondationnelles : un éco-système d'extensions du CIDOC CRM pour les sciences humaines et sociales ». *Actes des journées humanités numériques et Web sémantique* (Nancy, France). Sous la dir. de Nicolas Lasolle, Olivier Bruneau et Jean Lieber, p. 2-22. DOI : 10.5281/zenodo.7014341

Préface

Ce volume constitue les actes des journées d'étude « humanités numériques et Web sémantique », qui se sont déroulées les 21 et 22 juin 2021 au LORIA et le 21 juin 2022 aux Archives Henri-Poincaré, deux laboratoires situés dans l'agglomération de Nancy, France. Cette manifestation a ressemblé des chercheurs issus de disciplines en sciences humaines et sociales et des chercheurs menant des travaux relevant du Web sémantique. Les intervenants ont abordé des problématiques relatives à la définition d'ontologies de domaine, à la création d'outils pour la visualisation et l'exploration de corpus d'humanités numériques ainsi qu'à l'intégration de technologies du Web sémantique pour mener des raisonnements. De plus, les participants se sont intéressés aux évolutions des pratiques de recherche induites par les possibilités des outils numériques en croisant différents retours d'expérience. Ces discussions ont notamment permis de mettre en avant des aspects méthodologiques relatifs à l'utilisation et au développement d'outils numériques dans des contextes pluridisciplinaires.

11 articles ont été soumis et tous ont été évalués par deux membres du comité de lecture. 9 ont été retenus pour publication ; 7 sous la forme d'articles longs et 2 sous la forme d'articles courts.

Le site Web de la conférence est accessible à l'adresse suivante : https://humanumwebsem.sciencesconf.org/.

Cette manifestation a été organisée avec le soutien des laboratoires AHP-PReST et LORIA. Le comité d'organisation a bénéficié d'une aide de l'État, gérée par l'Agence Nationale de la Recherche, au titre du projet Investissements d'Avenir Lorraine Université d'Excellence, portant la référence ANR-15-IDEX-04-LUE.

Nicolas Lasolle, Olivier Bruneau et Jean Lieber, organisateurs de la journée « humanités numériques et Web sémantique ».

Interopérabilité des données de la recherche et ontologies fondationnelles : un écosystème d'extensions du CIDOC CRM pour les sciences humaines et sociales

Research data interoperability and foundational ontologies: an ecosystem of CIDOC CRM extensions for the humanities and social sciences

Francesco Beretta[1]

[1]*Laboratoire de recherche historique Rhône-Alpes, CNRS / Université de Lyon*

Abstract

Given the challenge posed by the giant knowledge graph established by big economic actors, that could virtually replace research in Humanities and Social Sciences (HSS) in order to respond to the public's concerns, the question arises as to how to enhance the value of research data through their publication and interconnection, in application of the FAIR principles. Both an epistemological and a semantic analysis show that the most relevant part of research data is factual information understood as a representation of the objects observed by the scientific disciplines, their properties and their relationships. This rich universe of information becomes comprehensible, and therefore reusable, through the application of foundational ontologies and a methodology based on the distinction between different levels of abstraction allowing the collective development of one or more shared and reusable domain ontologies. This vision will be built around the CIDOC CRM and the Semantic Data for Humanities and Social Sciences (SDHSS) high-level extension, as well as an ecosystem of published sub-domain extensions that can be easily managed through the ontome.net application. This will result in an interoperability that is semantically richer than the simple "technical" alignment of ontologies and less costly in terms of resources, and above all adapted to the scientific and humanistic project of the HSS.

Keywords

humanities and social sciences, research data, foundational analysis, semantic interoperability, abstraction layers, OntoME

1. Introduction

Le développement depuis une vingtaine d'années du principe des *linked open data* (LOD), ainsi que des méthodologies et technologies du web sémantique, a permis la mise en place de *knowledge graphs*, graphes du savoir [1], qui expriment les propriétés et les relations d'une multitude d'entités. La création de fichiers interconnectés d'autorités, tel les IdRef [2] ou le VIAF [3],

. *Workshop on Digital Humanities and Semantic Web*
. francesco.beretta@cnrs.fr (F. Beretta)
. http://larhra.ish-lyon.cnrs.fr/membre/76 (F. Beretta)

1. https://en.wikipedia.org/wiki/Knowledge_graph
2. https://www.idref.fr/
3. https://viaf.org/

ou de gazetteers tels Geonames [4] et ceux du réseau Pelagios [5], favorisent l'intégration de silos de données jusque là isolés grâce à l'identification et à la mise en relation de personnes, organisations, lieux, concepts, etc. Le web sémantique rend accessibles ces ressources, et leurs propriétés, sous forme d'informations dont le sens est explicité et formalisé par les ontologies afin d'être mobilisé tant par les humains que par les ordinateurs grâce aux technologies de *semantic reasoning* ou de *machine learning* [1]. Des ressources telles *data.bnf.fr* ou *scienceplus.abes.fr* rendent accessibles sous forme de données les notices bibliographique et un riche univers de métadonnées. Le potentiel de cette évolution a été reconnu par les moteurs de recherche qui améliorent la précision de leur résultats grâce à un artefact réalisé au cours des dernières années, qualifié de *giant knowledge graph*. Grâce aux progrès des technologies informatiques, et notamment l'extraction automatisée d'informations de textes, il est désormais possible d'envisager une alimentation du graphe du savoir rapide et quasi illimitée. Le graphe géant de Google comportait en 2020 cinq milliards d'entités et 500 milliards de « faits » [6].

Le potentiel de cette évolution ne peut laisser indifférents les chercheurs en sciences humaines et sociales (SHS) car ces méthodes et technologies vont non seulement impacter la production du savoir mais encore se substituer aux SHS en tant que fournisseur de réponses concernant les questions qui préoccupent la société civile et le public. En s'appropriant ces méthodologies, les SHS peuvent réagir au moins dans deux secteurs. Premièrement, c'est grâce à elles que tout le potentiel des principes FAIR, « make data Findable, Accessible, Interoperable, and Reusable » [7] va pouvoir se réaliser. Ces principes, formulés par un groupe de scientifiques issus du domaine des sciences naturelles et experts en sciences de l'information, ont pour finalité de promouvoir la réutilisation des données produites par la recherche afin de répondre à de nouveaux questionnements [8]. Les chercheurs sont ainsi invités à publier non seulement les résultats de leur enquêtes — le savoir produit — mais encore à mettre à disposition les données ayant servi à les établir [9]. Le jour où les données publiées par les chercheurs en SHS seront produites ou du moins mises à disposition dans les formats des LOD, et exprimées selon une ontologie standardisée, on réalisera pleinement les principes FAIR et on pourra construire un ou plusieurs *giant knowledge graphs* disciplinaires basé sur le capital-information cumulé par la recherche.

Deuxièmement, étant donné l'importance qui revient aux textes dans plusieurs disciplines SHS — la bibliographie et ses contenus, les sources comme traces du passé, les enquêtes comme reflet des opinions d'un groupe social, etc. — l'application aux documents écrits de méthodologies d'extraction automatisée de données structurées permettra d'enrichir considérablement les graphes de l'information et de rendre interrogeable et « actionnable » d'une manière totalement nouvelle le contenu des textes en révolutionnant la manière de produire le savoir. En d'autres termes, un changement de paradigme, c'est-à-dire une transformation des méthodes de

4. https://www.geonames.org/

5. https://pelagios.org/

6. https://en.wikipedia.org/wiki/Google_Knowledge_Graph

7. https://www.ccsd.cnrs.fr/principes-fair/. Cf. les instructions dans le cadre du Programme H2020 : Guidelines on FAIR Data Management in Horizon 2020, Version 3.0, 26 juillet 2016, de même que le site https://www.force11.org/group/fairgroup/fairprinciples.

8. « There is an urgent need to improve the infrastructure supporting the reuse of scholarly data » [2, 3]

9. Voir par exemple la revue Scientific data publiée par le groupe Nature : https://www.nature.com/sdata/, ou le Journal of Open Humanities Data : https://openhumanitiesdata.metajnl.com/

production du savoir et d'apprentissage de l'outillage disciplinaire, est en cours [10].

La condition pour la réalisation de ce projet est l'adoption par les communautés disciplinaires en SHS d'ontologies et de vocabulaires contrôlés à la fois standardisés, modulaires et extensibles, permettant de disposer d'une sémantique partagée clairement définie et flexible dans son application. Il importe en effet que l'identité des objets du discours scientifique, ainsi que le sens de leurs propriétés et relations, soient clairement explicités selon une méthodologie suffisamment robuste pour permettre aux données de répondre à la fois aux questionnements précis des chercheurs qui les ont produites et, plus tard, d'être réutilisées dans le contexte de nouvelles recherches, avec de nouvelles problématiques. L'enjeu est donc à la fois sémantique et épistémologique.

Dans la perspective d'une réflexion concernant l'impact de cette évolution sur la méthodologie scientifique en SHS, il faut s'interroger avant tout sur le contenu des données à partager, ainsi que sur la pertinence du terme *knowledge graph*. Une distinction importante s'impose en effet en SHS entre information et savoir : l'information peut être définie comme représentation de la réalité, le savoir comme interprétation de la réalité, compréhension de phénomènes complexes, de leur causes, de leur évolution probable. Certes les méthodologies sémantiques, les ontologies formelles, permettent de déduire de nouvelles informations à partir de celles dont on dispose, ce qui a amené à appeler ces objets des graphes de savoir. Mais du point de vue des SHS il ne s'agit pas d'un savoir au sens propre car celui-ci demande, au départ, la définition d'une problématique précise, d'un projet de recherche assorti d'un questionnement, et, à l'arrivée, la création dans l'esprit des chercheurs d'un modèle de la réalité, quantitatif ou qualitatif, qui sera partagé avec une communauté scientifique afin d'être discuté et révisé. Ce modèle sera proposé comme la meilleure explication disponible, jusqu'à nouvel avis, des structures, dynamiques, causes et évolutions possibles du monde humain et social, passé ou présent.

Dans cette contribution, je développerai tout d'abord ce dernier point, en précisant la distinction au point de vue épistémologique entre information et savoir, et entre information et données, telle qu'elle s'applique au sein du cycle de la connaissance en sciences historiques, et plus largement en SHS. Une définition précise de ces termes est indispensable afin de mettre en évidence l'enjeu central de l'application des ontologies dans ce domaine : c'est en effet l'information en tant que représentation du monde et des phénomènes humains qu'il convient de placer au cœur de l'interopérabilité des données et du graphe du web sémantique.

La deuxième partie sera dédiée à une présentation de la méthodologie proposée pour construire collectivement une conceptualisation à la fois clairement définie, extensible et suffisamment flexible pour être appliquée à la modélisation de l'information dans différents domaines des SHS. Au vu de la diversité de l'information mobilisée par les différentes disciplines il est impensable de disposer d'une seule ontologie couvrant tous les domaines : un dialogue intense est donc nécessaire entre les conceptualisations locales, imaginées par des projets précis, et une vision plus abstraite fondée sur les considérations et méthodologies développées depuis quelques décennies dans le domaine de la recherche sur les ontologies fondationnelles [11] et les méthodologies sémantiques. Comme support à cette démarche, le LARHRA a développé

10. J'ai formulé quelques réflexions à ce sujet dans un chapitre à paraître dans les actes des Premières journées historiographiques corses (juillet 2021).

11. https://fr.wiktionary.org/wiki/ontologie_fondationnelle

un service en ligne, OntoME [12], visant à permettre de gérer et de soutenir le développement modulaire et collaboratif d'un écosystème d'ontologies adaptées aux besoins de la recherche en SHS.

Dans la troisième partie, je présenterai les premiers résultats de ce processus de construction d'une conceptualisation susceptible de permettre l'interopérabilité de l'information. Il s'agira d'abord de proposer une analyse fondationnelle du CIDOC CRM, une ontologie formelle standardisée (ISO 21127:2014) et de plus en plus adoptée dans le domaine des SHS, conçue en vue de l'intégration de l'information issue des musées et de la conservation des biens culturels. Seront ainsi mis en évidence les atouts et les limites de cette ontologie au point de vue des SHS, tout en proposant une extension de haut niveau, formulée dans le projet *Semantic Data for Humanities and Social Sciences* (SDHSS), dont la finalité est de favoriser l'intégration des modèle conceptuels en cours de développement dans plusieurs projets au sein d'un écosystème d'ontologies permettant l'interopérabilité des données de la recherche.

2. Le cycle de la connaissance en sciences historiques

Données, information, savoir sont des termes polysémiques qu'il est important de définir avec précision. Je le ferai à l'aide de deux schémas qui résument le processus de connaissance en sciences historiques à partir de deux points de vue différents. Cette réflexion épistémologique modélise la pratique disciplinaire historienne mais elle est suffisamment générique pour pouvoir s'appliquer, moyennant les adaptations nécessaires, aux autres domaines de recherche en SHS. Le premier schéma s'inspire des étapes de l'élaboration du savoir formulées par Henri-Irénée Marrou sous forme de courbe parabolique dans un travail classique consacré au « métier d'historien » [4, p. 1502] (fig. 1). Le choix de présenter ici ce processus sous forme de cycle souligne la dimension itérative de la connaissance qui est propre à la démarche scientifique et qui s'applique également à la formulation et vérification (ou falsification) d'hypothèses propre aux sciences sociales [13]. Le deuxième schéma interprète du point de vue des sciences historiques la pyramide « données, information, savoir » utilisée par les sciences de l'information pour distinguer les différents niveaux de la connaissance [7] (fig. 2). La *connaissance* est entendue ici comme processus, le *savoir* comme contenu et résultat.

Comme le montre le schéma du cycle de la connaissance (fig. 1), toute recherche doit partir de la construction d'une problématique qui s'inscrit dans l'horizon du savoir existant, exprimé dans la bibliographie, et qui définit l'angle d'approche d'un sujet d'étude, la méthodologie et la question générale. Par exemple, dans une approche d'histoire intellectuelle des sciences, on peut s'interroger sur les conditions et les dynamiques de diffusion de l'héliocentrisme à l'époque moderne. Cette question générale doit être articulée dans un questionnement plus précis, concernant par exemple les carrières des astronomes et leur insertion dans les réseaux savants, en articulation avec l'analyse du contenu de leurs écrits, en restreignant éventuellement l'étude à une région ou à une catégorie spécifique. Cette première étape est indispensable afin de pouvoir ensuite choisir les sources à utiliser, ou les enquêtes à effectuer, et définir l'information qu'il faudra réunir pour répondre au questionnement.

12. https://ontome.net/
13. https://en.wikipedia.org/wiki/Scientific_method, [5, 6]

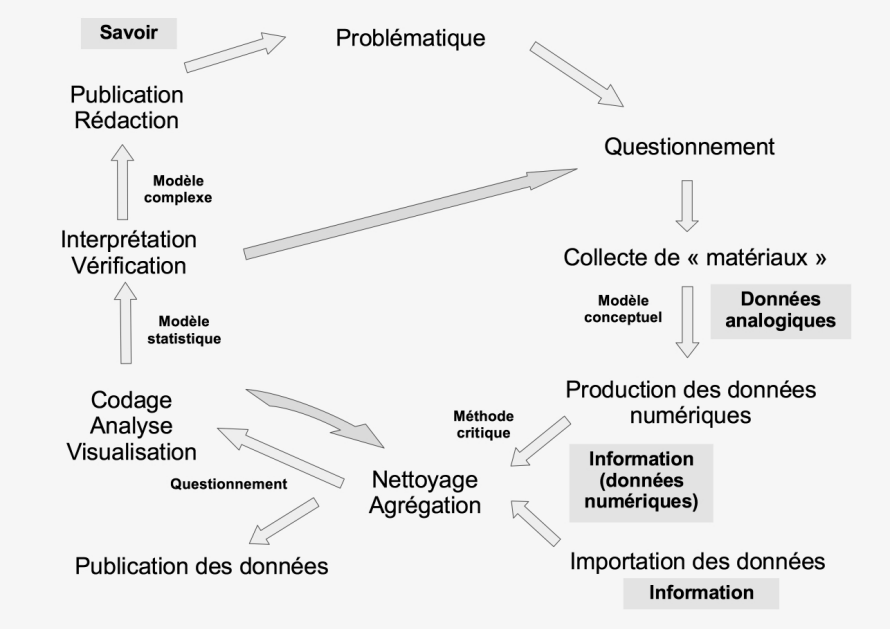


Figure 1 : Cycle de la connaissance en sciences historiques.

Nous nous situons à ce stade de la recherche au socle de la pyramide (fig. 2). Il importe de relever que les données sont à entendre ici au sens premier et étymologique, issu du latin *datum*, c'est-à-dire ce qui est donné et donc reçu par l'observateur comme état de fait, et non au sens de données numériques. On entend donc par données la réalité comme telle, dans son indépendance par rapport à l'observateur. À partir du questionnement les chercheurs en SHS doivent opérer un choix dans la masse que représentent les sources, ou toute autre trace disponible et/ou construite expérimentalement des activités humaines, afin de réunir l'information qui sera analysée et servira de fondement au savoir. Le questionnement permet de décider quelle information sera retenue systématiquement, et comment elle sera conceptualisée et produite. Se pose alors la question du modèle conceptuel et du choix de la technologie de stockage numérique car si une feuille de tableur peut faire l'affaire si on se limite à collecter systématiquement un certain nombre de propriétés d'une population d'individus de même type, dès qu'on souhaite renseigner des relations complexes entre différents objets, en lien avec l'espace et le temps, il est indispensable d'utiliser une base de données relationnelle ou orientée graphe afin de saisir toute la richesse de l'information.

Relevons quelques premiers acquis de cette analyse. L'information se situe au centre de la démarche scientifique. Elle peut être définie comme *représentation* de la réalité, et plus précisément comme représentation des objets du monde (les personnes, les organisations, les artefacts, etc.), de leur propriétés (les caractéristiques physiques des objets, les hobbys et les classes de revenus des personnes, les opinions, etc.) et de leur relations dans le temps et dans l'espace (les appartenances aux organisations, les échanges de messages ou de biens, les déplacements, etc.). Même si elle est conçue dans une perspective de *représentation*, donc avec une volonté explicite d'objectivité dans sa production, l'information est toujours construite,

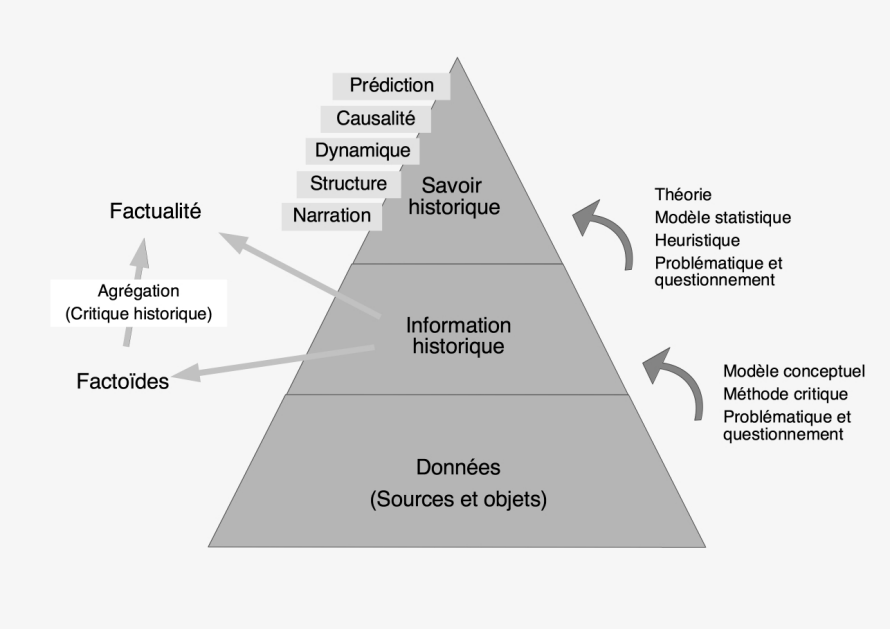


Figure 2 : Pyramide « sources, information, savoir » en sciences historiques.

elle résulte toujours d'un questionnement ou d'un point de vue. Par conséquent, les données de la recherche, telles que le contenu d'un tableur ou d'une base de données, ne sont point « données » au sens premier, ne représentent pas de manière immédiate la factualité car elles présupposent toujours un questionnement et une conceptualisation spécifiques qui ont permis leur collecte. Il est donc essentiel d'expliciter le contenu sémantique des données numériques et la modalité de leur production comme condition indispensable à leur réutilisation.

Notons aussi que, dans la pyramide, l'information s'articule à deux niveaux : on peut viser une reproduction fidèle du contenu des sources, ou une observation quotidienne des transactions économiques ou des manifestations des relations sociales contemporaines, en se situant à un niveau épistémologique qu'on appelle depuis quelques années celui des *factoïdes* [8]. Ce faisant on disposera d'une information volumineuse mais redondante, voire contradictoire au sujet des mêmes propriétés des objets. Prise comme telle cette information va inévitablement biaiser les résultats des analyses. Il faudra donc procéder à une agrégation des données (grâce à la méthode critique en histoire, ou à des méthodologies d'alignement plus ou moins automatisées, basées sur des modèles d'agrégation) afin d'obtenir des graphes d'information reproduisant autant fidèlement que possible la *factualité* des propriétés et des relations des objets étudiés. Disposer d'une information factuelle de qualité représente le socle indispensable de la production du savoir.

Une fois cette agrégation effectuée et l'information déposée sur un support numérique adéquat, il faudra la coder et la simplifier en fonction de la problématique de la recherche. C'est au niveau de cette deuxième opération d'agrégation qu'on injecte le questionnement afin de pouvoir appliquer à l'information, opportunément préparée, uniformisée et codée, une panoplie d'outils, logiciels statistiques, d'analyse de réseaux, de représentation et d'analyse

spatiale, etc. (fig. 1). Le modèle, au sens statistique, qui ressort de ces analyses a une fonction éminemment heuristique car les représentations mathématiques et visuelles qui résultent des outils logiciels nécessitent toujours une analyse critique ainsi qu'une contextualisation et une interprétation. En même temps, les logiciels d'analyse permettent de rendre visibles des phénomènes significatifs qu'il serait autrement impossible de voir «à l'œil nu»—par exemple la comparaison de segments de carrières et l'identification de profils prosopographiques récurrents chez des centaines d'astronomes à travers plusieurs siècles, en relation avec leur distribution dans l'espace géographique— et ce en dépit d'un volume et d'une complexité considérable de l'information disponible, opportunément compactée et simplifiée grâce aux regroupements et codages qui interviennent dans cette phase d'analyse.

Au terme de ce processus, les chercheurs aboutissent à la production du savoir comme réponse à une problématique et font connaître les résultats de leur enquête dans les publications. Il apparaît clairement de ces deux schémas qu'une distinction essentielle subsiste entre le *savoir* ainsi obtenu et *l'information* sur laquelle il se fonde car les (hypo-)thèses auxquelles la connaissance aboutit, relevant de la description des dynamiques complexes des phénomènes, de leurs structures et de leur causes, comportent toujours une synthèse de l'information et une interprétation qui dépassent la simple représentation de la factualité. Il est donc essentiel, dans la logique de la science ouverte, de publier non seulement le savoir obtenu mais encore les données-mêmes de la recherche, c'est-à-dire l'information collectée et analysée, afin de faciliter la vérification des hypothèses avancées en les exposant à la « falsification » dans la logique d'une démarche scientifique reproductible [14].

Cette analyse montre tout le potentiel pour la connaissance en SHS des nouvelles méthodologies en gestion des systèmes d'information car on peut désormais dépasser considérablement le volume de données que peuvent collecter individuellement les chercheurs et accéder à des gisements d'information de plus en plus riches et volumineux. En même temps, deux principes méthodologiques se dégagent qui doivent être appliqués rigoureusement afin de permettre la réutilisation des données de la recherche. D'une part, l'information exprimée dans les données numériques doit être le plus possible conçue en tant que représentation de la réalité factuelle, en dehors de tout codage découlant d'une problématique. L'agrégation et la simplification qui précèdent l'analyse doivent donc intervenir seulement dans une deuxième phase, alors que le partage des données concernera principalement l'information collectée dans la première phase de la recherche. D'autre part, les données qu'on souhaite partager doivent être produites grâce à une sémantique clairement définie. De plus, le processus de leur production doit être documenté soigneusement afin de permettre à d'autres chercheurs d'identifier les éventuels biais introduits dans le modèle conceptuel.

3. Ontologies fondationnelles et méthodologie de gestion d'ontologie

Reste une question essentielle qui est sous-jacente au scepticisme souvent exprimé quant à la possibilité effective d'une réutilisation des données produites par les SHS pour de nouvelles

14. https://fr.wikipedia.org/wiki/Réfutabilité

recherches : si l'information est le produit –– comme nous l'avons montré –– d'une construction conceptuelle qui découle de l'application d'un questionnement et adopte un conceptualisation en lien avec la problématique, n'y a-t-il pas là un obstacle majeur et quasi structurel à la réutilisation des données ? Une représentation de la réalité factuelle par l'information est-elle vraiment possible, ou en tout cas exprimable sous forme de données interopérables ?

La réponse à cette question —positive— nous est fournie par plusieurs décennies de publications dans le domaine des ontologies fondationnelles et méthodologies d'ingénierie des connaissances. Comme l'écrit l'un des acteurs de cette discipline, Giancarlo Guizzardi, dans un article à la fois critique et stimulant, l'interopérabilité de l'information, et la réalisation des principes FAIR, est possible seulement à condition d'adopter « formal, shared and explicit representations of conceptualizations, or what the area of knowledge representation has conventionally called ontologies». Et cet auteur de préciser que ce n'est pas le fait d'exprimer le modèle conceptuel d'un projet particulier grâce à la logique formelle ou à l'Ontology Web Language (OWL) qui crée une ontologie, mais bien le fait d'opérer une analyse des aspects essentiels de la réalité telle l'identité des objets qui la composent, leur rapports, leur compositions et dépendances, et ce en adoptant une conceptualisation de haut niveau qui est transdisciplinaire et qui peut s'appliquer à plusieurs domaines du discours scientifique. Tel est le rôle des ontologies fondationnelles [9], domaine dans lequel Guizzardi est actif en tant que l'un des créateurs de l'ontologie *Unified Foundational Ontology* (UFO) [10].

Un récent numéro de la revue *Applied Ontology* illustre de manière fort instructive cette démarche [11]. Les auteurs des principales ontologies fondationnelles, *Basic Formal Ontology* (BFO), *Descriptive Ontology for Linguistic and Cognitive Engineering* (DOLCE), *A Top Level Ontology within Standards* (TUpper), qui composent la norme ISO 21838, ainsi que UFO et quelques autres, ont été invités à proposer du point de vue de leur méthodologie d'analyse ontologique une modélisation de quelques questions classiques en ingénierie des connaissances concernant la description des artefacts et de leurs composantes, la modélisation de changements des propriétés des objets, ou la représentation des modifications des situations sociales. Le but est de permettre aux ingénieurs sémantiques de comprendre les fondements philosophiques des différentes ontologies –– car elles se basent généralement sur une tradition philosophique bimillénaire, avec des accents différents –– ainsi que les spécificités de leurs conceptualisations afin de choisir celle qui semble le plus efficace en termes d'analyse fondationnelle du domaine concerné.

Parmi ces ontologies, DOLCE se présente comme particulièrement adaptée à la perspective des SHS et elle connaît une certaine diffusion dans ce domaine [12, 13]. Nous avons choisi de la retenir en tant que référence pour notre analyse fondationnelle même si d'autres ontologies — notamment UFO avec le module UFO-C [14] — apportent également des perspectives analytiques intéressantes sur la modélisation des phénomènes sociaux. DOLCE est une ontologie de particuliers, c'est-à-dire qu'elle vise non pas à identifier la substance métaphysique de la réalité mais « to make explicit already existing conceptualizations through the use of categories whose structure is influenced by natural language, the makeup of human cognition, and social practices». Cette ontologie se prête donc particulièrement bien à réaliser le programme de création d'une conceptualisation interopérable de l'information en SHS — présenté ci-dessus — car si la réalité est bien le référent, l'analyse porte sur la conceptualisation exprimée dans le discours scientifique, ce dernier étant construit et susceptible d'évoluer.

De plus, DOLCE a été complétée non seulement par quelques extensions qui modélisent les rôles et les artefacts, voire les aspects sociaux et cognitifs, mais surtout par l'ontologie complémentaire *Descriptions & Situations* (D&S), développée dans le même projet d'origine et qui a comme domaine la modélisation fondationnelle des différentes perspectives des agents sur les mêmes événements du monde [13]. La notion de *situation* est définie comme interprétation d'événements à partir d'une conceptualisation particulière, c'est-à-dire de représentations partagées par les agents et exprimées par une *description* qui attribue rôles et connotations spécifiques aux participants de l'événement. D&S a été intégrée à DOLCE pour produire l'ontologie DOLCE Lite Plus (DLP), que nous avons adoptée comme référence pour notre travail analytique, et qui a été également reformulée et simplifiée dans DOLCE Ultra Light (DUL). Cette approche a permis d'aller jusqu'à modéliser l'activité des communautés scientifiques dans une optique constructiviste [15], ce qui correspond parfaitement, à mon sens, à l'analyse épistémologique présentée ci-dessus et répond au défi de trouver un équilibre entre une conceptualisation transdisciplinaire de l'information (DOLCE) et les spécificités de chaque discipline (D&S) : les mêmes faits peuvent correspondre à des « situations » différentes, c'est-à-dire à des interprétations différentes selon les points de vue des différentes disciplines. Ces interprétations doivent toutefois intervenir seulement comme *surcouche* à la production d'information, dans la deuxième phase de la recherche qui agrège et code les données afin les analyser et répondre au questionnement, comme nous l'avons vu.

DOLCE propose donc une conceptualisation — valide du moins dans le contexte de notre civilisation — qui permet la transdisciplinarité dans la production de l'information. Il est à relever que cette conceptualisation a été réalisée en utilisant la méthodologie OntoClean, développée par Nicola Guarino et Emil Welty, dont la finalité est de formaliser l'analyse fondationnelle autour de catégories fondamentales au point de vue philosophique, telle essence (et rigidité), identité, unité et dépendance [16]. Par conséquent, si on utilise DOLCE pour analyser la conceptualisation de son propre domaine on est déjà sur la bonne voie en termes de définition d'une ontologie robuste et interopérable, en on évitera bien des biais de modélisation grâce à cette méthode.

DOLCE répartit les particuliers, c'est-à-dire les entités auxquelles se réfère le discours scientifique, en quatre classes distinctes et sans intersection : les endurants, les perdurants, les qualités et les abstraits. La différence essentielle entre endurants et perdurants est leur rapport avec le temps : les endurants préservent leur identité à travers le temps, même si leur propriétés évoluent ; les perdurants, qui se développent dans le temps, et avec le temps, sont à chaque instant seulement partiellement présents, bien que identifiables dans leur ensemble. Endurants et perdurants sont reliés par la relation de participation des premiers dans les derniers, par exemple la participation de personnes à une réunion ou à une bataille. Il y a ensuite une distinction entre objets dépendants et indépendants, car un trou dans une chemise n'existe pas sans celle-ci, ni une grotte sans la montagne (il s'agit de *features*), et que la matière qui compose une table (le bois, *amount of matter*) a une identité qui est différente de la table elle-même, celle-ci résultant de sa forme (*physical object*). Dans la sphère des objets conceptuels on a des objets mentaux et sociaux, et notamment les rôles et les collectifs, qui résultent de la notion de classification et sont analysés dans les extensions de DOLCE.

Deux autres classes permettent d'articuler clairement le discours humain. D'une part les qualités, c'est-à-dire les propriétés observables des endurants ou perdurants, et qui leur sont spécifiques. Il s'agit notamment de l'espace occupé comme propriété des objets physiques alors

que la temporalité est une propriété spécifique aux événements. Il est à noter que les qualités sont conçues dans DOLCE comme inhérentes aux objets : chaque chaise a sa propre couleur à un moment donné. Chaque instance de la qualité couleur aura donc sa propre valeur, c'est-à-dire elle occupera un point ou une « région » dans un espace de référence, ce qui est exprimé par la notion de *region* comme sous-classe de la classe *abstracts* de l'ontologie. Les abstraits sont des entités du discours qui, n'ayant pas de propriétés temporelles ou spatiales propres, ni le statut de qualités, se situent en dehors des entités observables et, peut-on ajouter, apparaissent comme le produit de conventions de la communauté de recherche — par exemple les mesures métriques — permettant de situer les valeurs des propriétés dans un espace de référence. D'autres ontologies situent ces « abstraits » comme sous-catégories d'artefacts. Du point de vue épistémologique, il importe surtout de relever la distinction bien visible dans DOLCE entre les phénomènes et les espaces abstraits de référence, par exemple les lieux géographiques et les coordonnées dans le référentiel WGS84 qui permettent de situer les lieux dans l'espace abstrait du géoïde terrestre.

Si on applique ces catégories à l'analyse de l'information en tant que représentation de la réalité factuelle, on retrouve dans ces quatre classes les éléments essentiels présentés précédemment : les objets représentés par l'information sont les *endurants* (personnes, artefacts, groupes, etc.), leurs propriétés sont exprimées par des qualités (couleur, poids, effectif, etc.) qui se situent dans les espaces de référence propres aux différentes disciplines, tandis que leur relations et leur évolution dans le temps sont capturées grâce à leur participation dans les *perdurants*. Quant à leur évolution dans l'espace physique elle est conceptualisée dans DOLCE comme qualité des endurants et elle n'est donc qu'indirecte pour les perdurants dont la projection dans l'espace physique correspond à celle des objets qui participent aux événements.

On dispose ainsi de l'outillage conceptuel nécessaire pour construire des ontologies de domaine interopérables. En effet, on aura remarqué que les catégories présentées sont indépendantes de théories scientifiques ou problématiques spécifiques. Par conséquent, si l'information factuelle est capturée en adoptant cette conceptualisation, elle permettra de reproduire sous forme de données les propriétés et relations des objets de la manière la plus objective possible, tout en laissant aux disciplines scientifiques la tâche d'expliquer et d'interpréter ces mêmes propriétés et relations. En termes de méthode, il est nécessaire à ce stade de développer une ontologie de domaine, c'est-à-dire une conceptualisation d'un domaine particulier du discours scientifique car les ontologies fondationnelles proposent uniquement des « conceptual handles » mais n'ont pas vocation à être utilisées directement [11]. On pourrait opérer ce processus directement à partir de son propre modèle de recherche, évalué à l'aune des catégories fondationnelles, ce qui permettrait déjà de s'inscrire dans une logique d'interopérabilité. En vue de permettre l'interopérabilité des données produites par les disciplines scientifiques, il apparaît toutefois bien plus judicieux de procéder avec une méthodologie « par couches d'abstraction » (fig. 3, partie de gauche).

Selon cette méthode, il s'agit d'adopter une ontologie de domaine de haut niveau, une *core ontology*, fournissant les classes et propriétés de base permettant de décrire les objets étudiés par la discipline en question. Cette conceptualisation doit être soumise à vérification à l'aide des classes d'une ontologie fondationnelle, afin d'en améliorer qualité et expressivité. Ensuite, on peut développer des extensions par sous-domaines dans la discipline, par ex. l'histoire économique ou sociale pour les sciences historiques, proposant classes et propriétés qui capturent l'information concernant ces relations. Et enfin on choisira parmi les classes et propriétés

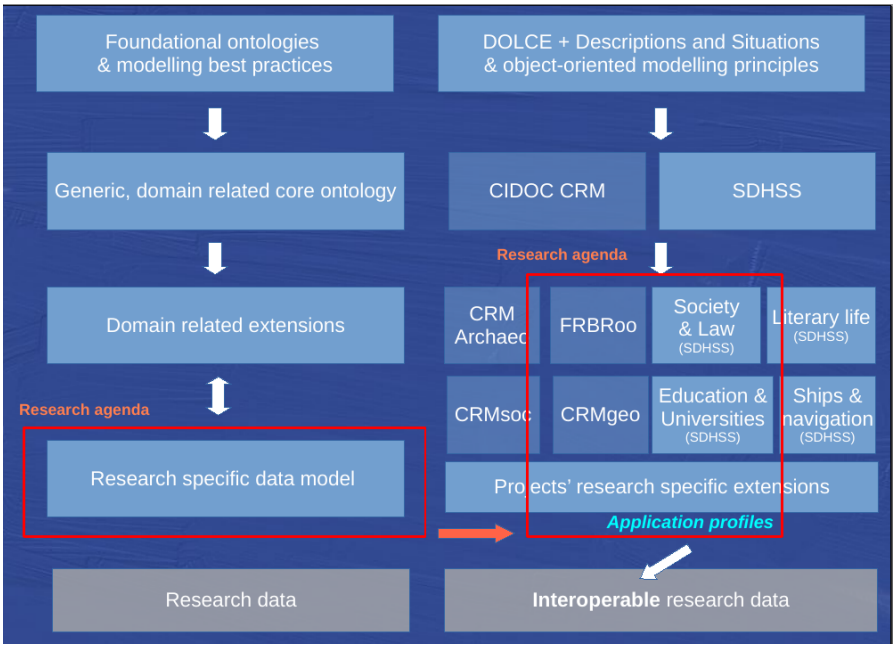


Figure 3 : Niveaux d'abstraction de la conceptualisation.

existantes celles à utiliser dans son propre projet, quitte à ajouter — si vraiment indispensable — celles qui manquent encore pour traiter l'information dont on dispose. L'avantage de cette méthode par couches d'abstraction est d'une part d'éviter de devoir réinventer à chaque projet une conceptualisation de domaine. D'autre part, la réutilisation de classes et propriétés existantes, clairement identifiées, facilite énormément l'interopérabilité. La condition est logiquement celle du respect strict de la conceptualisation de ces classes et propriétés, donc de la compréhension de leur « intension » , ce qui garantit l'interopérabilité grâce à une sémantique formalisée et partagée.

L'utilité de cette méthodologie par couches d'abstraction est apparue clairement dans l'évolution du projet *symogih.org* vers les méthodologies et technologies du web sémantique. J'illustrerai rapidement les étapes de ce processus car elles expliquent le choix de proposer le CIDOC CRM comme ontologie de domaine de haut niveau pour les SHS, de même que la nécessité d'ajouter une extension de haut niveau intégrant quelques aspects plus spécifiquement liés à l'information traitée par ces disciplines, notamment en lien avec la question de la conceptualisation des éléments essentiels de la vie sociale.

Le projet *symogih.org*, « Système modulaire de gestion de l'information historique » est né en 2008 de la volonté de quelques historiens du Laboratoire de recherche historique Rhône-Alpes (LARHRA) à Lyon de mutualiser les données structurées produites au cours de leur recherche afin de permettre leur réutilisation [17]. Par exemple, le projet SIPPAF –— financé par l'Agence nationale de la recherche entre 2007 et 2010 –— a abouti à la mise en place d'un système d'information prosopographique consacré au patronat français (XIX^{e}-XX^{e} siècles) [15].

15. http://www.patronsdefrance.fr/

Les informations produites au cours du projet continuent à être enrichies et utilisées plus de dix ans après la fin du financement, et ont notamment été réutilisées dans le cadre du projet SIPROJURIS consacré aux professeurs de droit en France de 1804 à 1950 [16]. Une cinquantaine d'autres projets, individuels ou collectifs, ont utilisé l'environnement virtuel de recherche (VRE) collaboratif mis en place par ce projet.

La réalisation de l'interopérabilité des informations dans le VRE a été réalisée grâce à l'application des deux principes : d'une part, nous avons soigneusement distingué entre la production des données en tant que représentation de la factualité et les classements qui précèdent l'analyse permettant de répondre au questionnement ; d'autre part, nous avons crée un modèle conceptuel générique et ouvert, suffisamment abstrait pour pouvoir s'adapter aux différents besoins de collecte d'information. Dans une logique de base de données générique, le modèle-même a été stocké sous forme de données, ce qui permet de le partager à l'intérieur du VRE et de le publier sur le site principal du projet. Le sens des données, i.e. la sémantique de l'information, est ainsi explicitée ce qui permet leur réutilisation [17]. En dépit du succès pratique de cette méthode, nous nous situions alors au niveau d'une simple conceptualisation de projet de recherche (cf. fig. 3) car aucune réflexion fondationnelle ni alignement sur une ontologie de référence n'avait été effectué.

La situation a commencé à évoluer dès 2013, dans une démarche de formalisation visant à adopter les technologies du web sémantique afin de mettre en relation les données du projet avec celles des autres fournisseurs, afin de s'inscrire dans la logique des LOD et des principes FAIR. Un premier processus de réécriture du modèle générique du projet symogih.org sous forme d'ontologie [18] a été abandonné car il a semblé bien plus judicieux, en termes d'interopérabilité, d'inscrire notre expérience de modélisation dans le contexte de la conceptualisation bien plus robuste et réfléchie proposée par le CIDOC CRM. Ce modèle conceptuel, ayant obtenu le statut de norme ISO en 2006, modélise le domaine des musées et présente donc des intersections importantes avec celui de la recherche historique. De plus, la méthodologie de développement du CRM, orientée objet et suivant des principes de conceptualisation proches d'OntoClean, fournit un système cohérent de classes de haut niveau à valeur universelle (comme les objets d'information ou les activités humaines), articulé dans une hiérarchie de classes avec héritage de propriétés construite autour d'une analyse fine des relations entre objets et événements [18]. En vertu de cette généricité, il a donc semblé judicieux de l'adopter comme *core ontology* pour le domaine des sciences historiques et, pourquoi pas, plus largement des SHS.

Nous avons donc entamé en 2016, lors d'un workshop à Héraklion, un processus d'alignement du modèle du projet symogih.org —avec ses 150 types d'information (classes et propriétés)— avec le CIDOC CRM, processus qui au fil du temps a montré la pertinence de ce choix mais aussi la difficulté d'aligner toute une partie de l'information déjà modélisée, et ce même en prenant en compte la famille d'extensions du CIDOC CRM, et notamment l'ontologie pour la description de la bibliographie et des sources FRBRoo [19]. En repensant à ce parcours aujourd'hui, et à la lumière des pages précédentes, les raisons de la difficulté rencontrée apparaissent clairement. D'une part, s'il y a certes intersection des domaines, il reste néanmoins une différence significative entre la

16. http://siprojuris.symogih.org/
17. http://symogih.org/?q=type-of-knowledge-unit-classes-tree
18. Cf. http://www.cidoc-crm.org/ .
19. http://www.cidoc-crm.org/collaborations

finalité du CIDOC CRM, c'est-à-dire l'intégration des données des musées grâce à un processus d'abstraction ontologique, et celle de la recherche historique qui s'efforce d'appliquer le principe fondamental de la production d'information comme représentation fine de la factualité, et qui demande par conséquent nuances et spécialisations. La mise en place d'une méthodologie par couche d'abstraction apparaît donc comme indispensable.

D'autre part, faisait alors défaut une analyse fondationnelle, en particulier grâce à l'application de la méthode OntoClean, permettant de mettre en évidence certains aspects non-compatibles des conceptualisations respectives au-delà d'une apparent homonymie des classes. L'adoption de DOLCE Lite Plus comme couche fondationnelle (fig. 3 partie de droite) a permis de clarifier le problème et d'individuer les aspects qui ne sont pas modélisés dans le CRM, ou et tous cas pas de manière entièrement satisfaisante [19]. Il est donc indispensable d'ajouter, au même niveau de *core ontology* du CRM, une extension que nous avons appelée *Semantic Data for Humanities and Social Sciences* (SDHSS) qui enrichit l'ontologie de haut niveau avec quelques classes indispensables pour structurer l'ensemble du domaine. Et aussi de compléter grâce à des extensions de niveau d'abstraction inférieure les lacunes dans les sous-domaines, comme la vie sociale et économique, ce qui ne peut se faire qu'en créant un écosystème d'extensions ayant vocation à être enrichi progressivement au fil des besoins des projets. Nous espérons que le développement de cet écosystème deviendra de plus en plus participatif afin de permettre à un grand nombre de projets en SHS de tester les conceptualisation proposées dans leur recherche et de construire progressivement une vraie interopérabilité sémantique des données.

Cet objectif explique la création d'un support en ligne indispensable : l'application de gestion collaborative d'ontologies mise en place par le LARHRA à partir de 2017, OntoME (*Ontology Management Environment*) [20], dont une première phase de développement vient d'aboutir et qui a été adoptée par différents projets [21]. OntoME permet de gérer de multiples espaces de noms, disposant de gestion de droits autonomes par projet, d'importer et d'exporter des ontologies en RDFS et OWL-DL, de créer des profils applicatifs à utiliser dans des VRE de production des données, telle *geovistory.org*, Wisski ou autres. OntoME permet aussi de créer des extensions de bas niveau, telles celles du projet *Maritime History* [20] ou de l'ANR *Processetti* [22], adaptées aux besoins de production d'information des recherches respectives mais développées à partir de la méthodologie par couches d'abstraction présentée ci-dessus. Le cycle de vie de ces extensions peut se limiter à la durée du projet, ou alors elles pourront être réutilisées et complétées par de nouveaux projets travaillant sur les mêmes sous-domaines, dans la logique d'un écosystème dynamique et évolutif.

La promotion de cette vision d'intégration sémantique des données a également motivé la création du consortium Data for History [23], constitué en novembre 2017 lors d'un workshop organisé à l'École normale supérieure de Lyon, suivi par un deuxième atelier lyonnais en 2018, puis par une rencontre à Leipzig en 2019 [24] et par la première conférence internationale (en ligne) en mai-juin 2021 organisée par la chaire d'histoire numérique de l'Université Humboldt

20. https://ontome.net/

21. En particulier, deux projets financés sur fonds européens ont utilisé OntoME pour la préparation du modèle de données : Silknow (projet ERC) et Read-it (projet JPICH).

22. https://ontome.net/profile/15

23. http://dataforhistory.org/

24. http://dataforhistory.org/3rd-data-for-history

de Berlin [25], et qui se prolonge actuellement par les Data for History Lectures [26].

4. Une extension de haut niveau du CIDOC CRM : SDHSS

Dans les pages qui précèdent j'ai indiqué les raisons qui amènent à adopter le CIDOC CRM (désormais CRM) comme ontologie de domaine de haut-niveau pour la modélisation de l'information en sciences historiques, et plus largement en SHS, mais en même temps la nécessité d'étendre cette ontologie avec les classes qui manquent à ce même niveaux d'abstraction, pour répondre aux besoins de la recherche. La finalité de cette démarche est d'arriver à exprimer l'information produite au cours de la recherche en tant que représentation d'objets, de leurs propriétés et de leur relations, avec le plus possible d'objectivité et de rigueur. La question de la manière de conceptualiser la production de l'information, et d'exprimer sa qualité, pourtant également essentielle en vue de l'interopérabilité ne sera pas abordée ici, d'autant plus qu'elle a déjà donné lieu à un certain nombre de proposition de solution, par exemple l'ontologie *Historical Context Ontology* (HiCO) [27], extension de PROV-O [28].

Le projet envisagé doit partir d'une analyse du CRM à l'aune de la méthodologie OntoClean et des ontologies fondationnelles, donc DOLCE dans notre cas. Cette étude a déjà été entreprise et quelques limites ou inconsistances du CRM ont été mises en évidence, avec des propositions d'amélioration de la formalisation de l'ontologie dont j'évoquerai quelques aspects dans les pages qui suivent [21]. On peut découvrir la structure de l'ontologie en inspectant l'arborescence des classes publiées dans OntoME [29]. En dépliant progressivement l'arbre et en parcourant ses branches, on trouvera les classes que je présenterai et on pourra accéder à la définition de leur « intension » dans les *scope notes*, et celles de leurs propriétés. À noter que l'environnement en ligne OntoME est fondamentalement agnostique, on peut y héberger toute ontologie entendue au sens de modèle du monde (et non de vocabulaire contrôlé ou de collection d'instances). Nous avons toutefois souhaité, dans cette phase, promouvoir « en première page » la vision présentée ici : dans l'arborescence, sans besoin de se connecter, on trouvera ainsi outre le CRM et FRBRoo, les espaces de noms qui font partie du projet SDHSS. Afin de les distinguer, je les préfixerai avec crm pour le CRM et sdh pour l'extension de haut niveau [30].

La classe racine, *crm:E1 Entity*, contient tous les objets du domaine de discours du CRM. On remarquera que les valeurs, les *literal values* au sens du RDF, n'en font pas partie et sont réunies dans la classe *crm:E59 Primitive Value*. Elles se situent donc en dehors de l'ontologie qui renvoie aux standards existants pour exprimer ces valeurs. Si on déplie l'arbre, on remarque les deux classes essentielles *crm:E77 Persistent Item* et *crm:E2 Temporal Entity*, correspondant respectivement aux classes *Endurant* et *Perdurant* de DOLCE. Manquent en revanche les classes *Quality* et *Abstract*, alors qu'il y a quatre autres classes de niveau racine (*crm:E54 Dimension, crm:E53 Place, crm:E52 Time Span, crm:E92 Spacetime Volume*). Elles se présentent, à la lumière

25. https://d4h2020.sciencesconf.org/
26. http://dataforhistory.org/news
27. https://marilenadaquino.github.io/hico/
28. https://www.w3.org/TR/prov-o/
29. https://ontome.net/classes-tree
30. La version du CRM utilisée est la 6.2, en ligne au moment de la rédaction de cette contribution. Elle sera prochainement remplacée par la nouvelle version 7.1.1, candidate à la nouvelle version de la norme ISO

de la conceptualisation de DOLCE, comme des régions, des sous-classes d'*Abstract*, car elles correspondent à une position particulière dans un espace de référence conventionnel. Elles sont donc réunies dans la classe *sdh:C5 Region* de l'extension afin de souligner clairement cette analyse et d'éviter les confusions.

Notons à ce sujet qu'on assiste fréquemment à la méprise de projets qui utilisent la classe *crm:E53 Place* pour modéliser les lieux géographiques : selon le CRM, *place* est une pure étendue dans un espace de référence, et mériterait donc plutôt de s'appeler *space*, ce qui est confirmé par le fait que selon le CRM on peut prendre en photo une instance de *crm:E27 Site* — généralement utilisé pour les sites archéologiques mais en fait un lieu géographique — mais pas une instance de *crm:E53 Place* dont la substance est purement géométrique [31]. La classe *sdh:C13 GeographicalPlace* a donc été ajoutée dans l'extension afin de clarifier l'identification de l'objet « lieu géographique » et de rendre compte du fait qu'un lieu peut se trouver projeté, au cours du temps, dans différentes instances de *crm:E53* Place, telle une ville ou un territoire dont les surfaces évoluent avec les années.

Concernant la classe *crm:E77 Persistent Item* et ses sous-classes, elles expriment une conceptualisation pas très éloignée, à première vue, de celle de DOLCE, et comportent des objets indépendants et des *features* qui leur sont associées, des objets physiques et leur pendant non-matériel. Il y a toutefois quelques spécificités qui ont été mises en évidence car non-conformes à la méthode OntoClean. Tout d'abord une distinction entre agent (*crm:E29 Acteur*) et objet « inerte » (*crm:E70 Thing*) qui se fonde davantage sur l'intentionnalité que sur un classement plus objectif, les acteurs étant les personnes, "individually or in groups, who have the potential to perform intentional actions". Les animaux et les agents non-humains paraissent donc exclus et se retrouvent virtuellement sous la forme de *crm:E24 Physical Man-Made Thing ou crm:E20 Biological Object*, plus bas dans la hiérarchie, mais on est surpris alors de rencontrer de nouveau, à ce échelon de la taxonomie, les personnes, ici comprises dans leur matérialité biologique, ou « animalité ». La taxonomie de DOLCE est bien plus stricte au point de vue de la méthode OntoClean.

Cette impression de « flou » ontologique apparaît aussi dans la définition de la classe *crm:E72 Legal Object*, distincte en apparence dans l'arborescence de la classe *crm:E71 Man-Made Thing*, bien qu'en réalité la classe *crm:E24 Physical Man-Made Thing* apparaisse plus bas dans la hiérarchie en tant que sous-classe des deux classes précédentes. La fonction de la classe *crm:E72 Legal Object* est de regrouper les objets sur lesquels un droit appartenant aux acteurs peut s'exercer. Il a été remarqué à juste titre que cette classe est donc *anti-rigid* au sens d'OntoClean, c'est-à-dire que le fait d'être soumis à propriété ou autre droit est certes possible, mais non essentiel à la définition de la classe, ce qui inviterait à enlever *crm:E72 Legal Object* de la hiérarchie des classes et à exprimer cette connotation légale avec une autre conceptualisation.

Une remarque méthodologique importante s'impose à ce stade de la discussion. Même si le CRM a été développé en appliquant une analyse précise de l'identité, unité et essence des classes, la méthodologie qui explique les taxonomies n'est pas celle d'OntoClean mais bien une approche orientée objet qui se construit à partir de l'analyse des propriétés, entendues ici

31. Voire la scope note de la classe *crm:E27 Site* : «In contrast to the purely geometric notion of E53 Place, this class describes constellations of matter on the surface of the Earth or other celestial body, which can be represented by photographs, paintings and maps», https://ontome.net/class/26.

comme expression des relations entre entités. La fonction de la classe crm:E72 Legal Object est donc d'apporter à ses classes descendantes les propriétés qui associent ces entités aux acteurs exerçant un droit sur elles (*crm:P105 right held by*) ainsi qu'au droit exercé lui-même (*crm:P104 is subject to crm:E30 Right*), ce dernier étant exprimé sous forme d'objet propositionnel. Le CRM applique une approche d'héritage multiple qui combine au sein de la hiérarchie des classes aussi bien celles qui sont « essentielles » au sens de OntoClean que celles qui apportent des qualifications supplémentaires sous forme de propriétés, ce qui a amené à appeler le CRM une "property-centric ontology" [22]. Les propriétés sont à entendre ici au sens de relations, non de qualités.

Les raisons du choix de cette approche —qui combine deux méthodologies en apparence incompatibles— ont été exprimées clairement par son créateur, Martin Doerr, dans un article intitulé *The Dream of a Global Knowledge Network* qui non seulement présente le CRM comme "nearly generic information model" mais qui, sur la base de cette approche, ouvre la voie à la réalisation du projet d'interopérabilité et de scalabilité de la réutilisation de l'information que nous avons présenté dans l'introduction [23]. Il faut reconnaître à cet auteur, et aux experts dont il a su s'entourer, tout le mérite d'avoir adopté une méthode quelque peu « hybride » mais très efficace en termes de réalisation des objectifs d'interopérabilité envisagés. En même temps, une analyse ontologique fondationnelle permet d'identifier les parties à compléter ou à préciser dans le CRM, notamment si on souhaite l'utiliser dans le domaine du discours de la recherche en SHS.

Parmi les questions les plus significatives en termes de complément indispensable, retenons celle du traitement des propriétés des objets, entendues au sens de *Quality* de DOLCE. Notons préalablement que la notion de *crm:E2 Temporal Entity* recouvre tous les phénomènes qui se passent dans une période limitée de temps, avec une référence explicite à la notion de *Perdurant* de DOLCE. Une observation attentive de cette classe du CRM, dans une perspective *property-centric*, montre qu'en effet toutes ses propriétés expriment soit une relation temporelle avec d'autres phénomènes — au sens des propriétés temporelles d'Allen [24] — soit une relation à un *crm:E52 Time-Span* dont la fonction est d'établir une position spécifique dans le référentiel abstrait du temps. Notons aussi que, en dépit de l'identité d'appellation, l'essence ontologique de la classe *TemporalEntity* de la *Time Ontology* in OWL [32] n'est pas la même celle-ci correspondant en fait à *crm:E52 Time-Span*, car il s'agit bien d'une *Temporal Region* au sens de DOLCE, alors que *crm:E2 Temporal Entity* représente un phénomène susceptible d'être observé, voire photographié.

Parmi les sous-classes de *crm:E2 Temporal Entity* on compte *crm:E4 Period*, qui est la racine de la conceptualisation de tous les événements physiques ou culturels, ainsi que *crm:E3 Condition State* qui a été interprété au sens de phase mais qui pourrait en fait être également compris comme classe équivalente à *Quality* de DOLCE, dont l'absence dans le CRM a été relevée. La seule classe correspondante semble être *crm:E16 Measurement* qui utilise la classe *crm:E54 Dimension* afin de renseigner une région dans un espace abstrait quantitatif défini par une unité de mesure. Notons que le phénomène capturé par la classe *crm:E16 Measurement* est le moment de l'observation, par exemple celle de la longueur d'un pont à une date donnée. Cette classe se situe donc dans la perspective des factoïdes car on pourrait renseigner de multiples fois, dans le système d'information, la même longueur que ce pont mesurait à des moments différents,

32. https://www.w3.org/TR/owl-time/#time:TemporalEntity

alors que l'information agrégée dont on souhaiterait disposer aux fins de la recherche est que tel pont avait comme qualité telle longueur à une époque donnée avant d'être transformé avec une longueur différente en telle année, ce qui représente une information factuelle agrégée.

Il semble donc judicieux d'ajouter dans l'extension la classe *sdh:C1 Entity Quality* qui correspond à la notion de qualité de DOLCE et permet d'ajouter une composante essentielle dans la conceptualisation de la recherche en SHS. On pourra en effet traiter des qualités tant qualitatives que quantitatives, et leur évolution dans le temps, indépendamment et en complémentarité par rapport aux événements qui structurent le CRM. La définition de *sdh:C1 Entity Quality*, créée en tant que sous-classe de *crm:E2 Temporal Entity*, s'exlique par la méthodologie de modélisation « hybride » discutée ci-dessus. Car si cette classe correspond, d'une part, à une *time-indexed quality* au sens de DOLCE elle est, d'autre part, déclarée comme sous-classe de *crm:E2 Temporal Entity*. Elle se présente donc, dans son essence, comme articulant un phénomène observable, limité dans le temps, et en même temps une qualité inséparable de l'objet dont elle représente une propriété qualitative ou quantitative. Deux propriétés, *sdh:P8 effects* et *sdh:P9 ends*, associent les événements du monde physique aux qualités.

Si aucune propriété n'associe directement cette classe de haut-niveau à l'ensemble des objets du CRM c'est que les qualités ne sont pas les mêmes pour l'ensemble des entités, et qu'il est donc plus opportun d'introduire des sous-classes exprimant la relation de différentes qualités avec différents types d'objets. Dans la perspective de DOLCE, cette classe inscrit donc *Quality* comme sous-classe de *Perdurant*, en principe disjointes ! Grâce à cette entorse à la méthode OntoClean —car l'essence de cette qualité est très englobante et donc nécessairement imprécise, et de surcroît fusionnée avec la notion de perdurant— cet artefact ontologique se présente en revanche comme une composante puissante de l'extension car elle permet de conceptualiser bon nombre de propriétés des objets qui apparaissent comme des phénomènes limités dans le temps et qui, comme tels, sont inexprimables dans l'approche « centrée événement » propre au CRM.

C'est le cas en particulier de la conceptualisation de la vie mentale et sociale qui est à la racine de la plupart des phénomènes étudiés par les SHS. Le CRM restreint son analyse de la vie mentale des humains à ce qui est exprimé dans la matérialité: "What goes on in our minds or is produced by our minds is also regarded as part of the material reality, as it becomes materially evident to other people at least by our utterances, behavior and products" [33]. Certes des classes existent, telle *crm:E66 Formation* ou *crm:E68 Dissolution*, permettant de traiter le début et la fin d'existence des groupes, ou *crm:E85 Joining* et *crm:E86 Leaving*, pour exprimer les rapports des acteurs avec les groupes. Mais ces classes sont conceptualisées en tant que projection dans le monde des événements physiques, d'une réalité intentionnelle qui classifie une personne comme étant membre d'un groupe. Comment traiter, à partir de cette approche, les rôles politiques des personnes, les sièges légaux des entreprises, en un mot : les propriétés complexes des objets qui résultent de phénomènes sociaux limités dans le temps et reconnus comme tels ?

L'extension SDHSS introduit la classe *sdh:C4 Intention* en tant que sous-classe de *sdh:C1 Entity Quality* afin d'intégrer l'intentionnalité tant dans le sens de la philosophie sociale que de la psychologie sociale et de la sociologie, autour de la notion de représentation(s), mentales ou collectives. Cette notion est conceptualisée en accord avec une compréhension généralisée dans

33. Definition of the CIDOC Conceptual Reference Model, Version 7.1.1, April 2021.

ces disciplines — formulée de manière particulièrement précise par le philosophe John Searle [25] — qui observent que les humains, individuellement ou en groupe, portent leur attention sur les objets à travers leurs représentations [34]. Dans la logique de l'approche épistémologie présentée précédemment, la conceptualisation de la classe *sdh:C4 Intention* n'intervient donc pas dans le débat philosophique, ou dans l'explication scientifique de ce phénomène, mais se limite à construire un concept qui capture un phénomène observable — tel le concept de masse en physique — tout en laissant aux différentes disciplines scientifiques le soit de le définir et de l'expliquer.

L'intentionnalité est donc conçue comme une qualité propre à l'esprit d'une personne, ou de plusieurs personnes dans un logique d'intentionnalité collective, qui adhérent mentalement à des représentations portant sur un objet. La modélisation proposée s'abstient d'entrer dans le débat épistémologique et observe l'existence d'instances identifiées par la classe *sdh:C9 Intentional Entity* — qu'elles soient des humains pris individuellement ou en groupe, des animaux ou des artefacts digitaux — capables d'effectuer un classement concernant un objet du monde, à un moment du temps donné, une connotation dans le contexte de représentations exprimées par la classe *crm:E89 Propositional Object*. L'intentionnalité se présente ainsi comme une qualité du support matériel, biologique ou —si on veut— numérique, individuel ou collectif, qui permet de rendre compte de phénomènes comme l'attribution de rôles aux personnes, la propriété d'objets, l'appartenance aux groupes, etc. dont la réalité n'est pas inhérente aux objets concernés (les personnes ou les objets) mais est conceptualisée comme une qualité des observateurs. On peut ainsi rendre compte dans l'ontologie du fait que dans le même pays, au même moment, deux groupes distincts d'observateurs considèrent telle personne comme élue légitimement ou non à la fonction de président.

Cette conceptualisation s'inspire et s'inscrit dans l'analyse ontologique de D&S autour de la classe *Situation* conçue comme interprétation spécifique, et virtuellement discordante, des mêmes événements du monde, conceptualisation qui a été développée dans une perspective constructiviste autour de la notion d'*intentional collective* [15]. La classe *sdh:C4 Intention* permet ainsi de capturer l'information produite par l'observation de phénomènes sociaux et devient la racine d'une multitude de sous-classes —dans différentes extensions de plus bas niveau d'abstraction qu'on ne peut présenter ici— en acquérant une position équivalente à la classe *crm:E5 Event*. La cohérence entre le niveau intentionnel et le niveau de la matérialité physique qui fonde le CRM ("material reality is regarded as whatever has substance that can be perceived with senses or instruments" [35]) est établie par la propriété *sdh:P43 has setting* qui associe le phénomène mental à son substrat situé dans le monde physique. Par exemple, les phénomènes intentionnels que provoque la lecture de ce texte dans l'esprit du lecteur se réalisent par le fait que ses yeux parcourent les signes et que ses neurones les interprètent, et ce qu'il soit assis, debout ou qu'il marche, ou les trois successivement, à condition qu'il ait préalablement pris en main le support sur lequel se trouve cette instance de la classe *crm:E73 Information Object*. Ces phénomènes sont complémentaires et inséparables, mais distincts.

34. https://plato.stanford.edu/entries/intentionality/ ; https://plato.stanford.edu/entries/collective-intentionality/
35. Definition of the CIDOC Conceptual Reference Model, Version 7.1.1, April 2021.

5. Conclusion

Au terme de ce parcours d'analyse épistémologique et sémantique, il me semble important de retenir trois éléments. Premièrement, parler d'interopérabilité des données de la recherche en SHS présuppose de s'interroger sur le contenu de celles-ci, tout en les situant dans le contexte d'une analyse de la production du savoir. Le contenu des données numériques le plus pertinent et utile aux fins de leur réutilisation en accord avec les principes FAIR consiste dans *l'information* entendue comme représentation des objets observés, de leurs propriétés et de leurs relations. Différentes disciplines et projets de recherche s'intéresseront à différents aspects de la réalité, à différents objets considérés à partir de différents angles de vue et problématiques. Toutefois, si on applique rigoureusement la séparation indispensable entre deux phases distinctes de la recherche, l'une produisant les données numériques comme véhicule d'une information la plus objective possible, l'autre introduisant les codages qui permettent l'analyse, on obtiendra un riche univers d'information réutilisable permettant de représenter différentes facettes de la réalité dans un graphe cumulatif de volume et de qualité de plus en plus importants.

Deuxièmement, ce projet ne peut être réalisé qu'à condition d'appliquer les méthodes établies d'analyse ontologique, notamment grâce à l'utilisation d'ontologies fondationnelles, et à la distinction de différents niveaux d'abstraction permettant de développer collectivement un écosystème d'ontologies partagées et réutilisables. L'application en ligne *ontome.net* a été conçue comme support permettant de faciliter la mise en œuvre de cette vision, afin d'offrir aux différents projets la possibilité d'adopter des modèles de données spécifiques à leur recherche tout en réutilisant le plus possible l'existant et en les inscrivant dans une ontologie articulée en différents niveaux d'abstraction.

Troisièmement, il semble judicieux d'adopter le CIDOC CRM, couplé avec le FRBRoo et autres extensions, pour disposer d'une *core ontology* mettant à disposition les classes de haut-niveau indispensables pour décrire une partie importante de l'information relevant du domaine des SHS. Mais il est en même temps indispensable de l'incrémenter avec une extension de haut niveau, *Semantic Data for Humanities and Social Sciences* (SDHSS), afin de couvrir l'ensemble du domaine et d'ajouter ensuite à des niveaux inférieurs d'abstraction les extensions de sous-domaine indispensables à la recherche. Cet écosystème cohérent d'ontologies permettra de mettre à disposition des SHS toute une série de conceptualisations réutilisables afin de garantir une interopérabilité bien plus riche sémantiquement que le simple alignement 'technique' d'ontologies et beaucoup moins coûteuse en temps et ressources que le fait de devoir réinventer une conceptualisation pour chaque projet.

Cette vision et cette démarche méthodologique visent à favoriser l'application des principes FAIR dans le domaine de la recherche en SHS et à permettre de réaliser un graphe géant de l'information au service de ces disciplines. Reste à savoir si les communautés de recherche sauront s'ouvrir à cette transition à la fois épistémologique et pratique. Pour réussir, elle demande une nouvelle forme d'engagement collectif dépassant les cloisons disciplinaires et les logiques de projet, imperméables à la vision des principes FAIR. Elle représente aussi un engagement citoyen des SHS, pour prendre position face au pouvoir économique et symbolique des géants du web basé notamment sur les graphes du savoir inaccessibles et orientés vers la rentabilité financière. Un graphe géant de l'information, maintenu collaborativement par la recherche en SHS, permettrait de défendre une analyse des réalités du monde à la fois critique et humaniste.

Références

[1] J. Dörpinghaus, A. Stefan, B. Schultz, M. Jacobs, Context mining and graph queries on giant biomedical knowledge graphs, Knowledge and Information Systems 64 (2022) 1239–1262.

[2] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, et al., The FAIR Guiding Principles for scientific data management and stewardship, Scientific data 3 (2016) 1–9.

[3] B. Mons, E. Schultes, F. Liu, A. Jacobsen, The FAIR principles : first generation implementation choices and challenges, 2020. doi :10.1162/dint_e_00023.

[4] H.-I. Marrou, Comment comprendre le métier d'historien, in : S. Charles (Ed.), L'histoire et ses méthodes, Paris, Éditions Gallimard, 1961, pp. 1465–1540.

[5] W. Little, R. McGivern, N. Kerins, Introduction to sociology, BCampus, 2016. URL : https://opentextbc.ca/introductiontosociology2ndedition/, 2nd Canadian edition, chapter 2. Consulté le 31.05.2022.

[6] R. S. Jhangiani, I. Chiang, P. C. Price, Developing a hypothesis, in : Research methods in psychology, BC Campus, 2015. URL : https://open.bccampus.ca/browse-our-collection/find-open-textbooks/?uuid=497a78e4-1384-4334-bcc2-e9040a436322, 2nd Canadian edition, chapter 10. Consulté le 31.05.2022.

[7] J. E. Rowley, The wisdom hierarchy : representations of the DIKW hierarchy, Journal of Information Science 33 (2007) 163–180. doi :10.3917/dunod.praxj.2019.01.

[8] M. Pasin, J. Bradley, Factoid-based prosopography and computer ontologies : towards an integrated approach, Literary and Linguistic Computing 30 (2015) 86–97.

[9] G. Guizzardi, Ontology, ontologies and the "I" of FAIR, Data Intelligence 2 (2020) 181–191. doi :10.1162/dint_a_00040.

[10] G. Guizzardi, A. Botti Benevides, C. M. Fonseca, D. Porello, J. P. A. Almeida, T. Prince Sales, UFO : Unified Foundational Ontology, Applied Ontology (2021) 1–44. doi :10.3233/AO-210256.

[11] S. Borgo, A. Galton, O. Kutz, Foundational ontologies in action, Applied ontology 17 (2022) 1–16.

[12] S. Borgo, C. Masolo, Foundational choices in DOLCE, in : S. Staab and R. Studer (Ed.), Handbook on ontologies, Springer-Verlag Berlin Heidelberg, 2009, pp. 361–381. doi :10.1007/978-3-540-92673-3_16.

[13] C. Masolo, S. Borgo, A. Gangemi, N. Guarino, A. Oltramari, Wonderweb deliverable D18-ontology library (final report), 2003. Laboratory for Applied Ontology, Trento.

[14] R. S. Guizzardi, G. Guizzardi, Ontology-Based Transformation Framework from Tropos to AORML., 2011.

[15] A. Gangemi, J. Lehmann, C. Catenacci, Norms and plans as unification criteria for social collectives, Autonomous Agents and Multi-Agent Systems 17 (2008) 70–112. doi :10.1007/s10458-008-9038-9.

[16] N. Guarino, C. A. Welty, An overview of OntoClean, Handbook on ontologies (2009) 151–171.

[17] F. Beretta, P. Vernus, Le projet SyMoGIH et la modélisation de l'information : une opération scientifique au service de l'histoire, Les Carnets du LARHRA (2012) 81–107.

[18] F. Beretta, L'interopérabilité des données historiques et la question du modèle : l'ontologie

du projet SyMoGIH, Presses universitaires de Paris Nanterre, 2017, pp. 87–117.
[19] F. Beretta, A challenge for historical research : making data FAIR using a collaborative ontology management environment (OntoME), Semantic Web 12 (2021) 279–294. doi :10.3233/SW-200416.
[20] F. Beretta, V. Alamercery, S. Derks, L. Petram, J. Schneider, Geohistorical FAIR data : data integration and Interoperability using the OntoME platform, in : Time Machine Conference 2019, 2019.
[21] E. M. Sanfilippo, B. Markhoff, P. Pittet, Ontological Analysis and Modularization of CIDOC-CRM, in : B. Brodaric, F. Neuhaus (Eds.), Formal Ontology in Information Systems : Proceedings of the 11th International Conference (FOIS 2020), volume 330, IOS Press, 2020, pp. 107–121. doi :10.3233/FAIA200664.
[22] M. Doerr, The CIDOC conceptual reference module : an ontological approach to semantic interoperability of metadata, AI magazine 24 (2003) 75–75. doi :10.1609/aimag.v24i3.1720.
[23] M. Doerr, D. Iorizzo, The dream of a global knowledge network—a new approach, Journal on Computing and Cultural Heritage (JOCCH) 1 (2008) 1–23. doi :10.1145/1367080.1367085.
[24] J. Holmen, C.-E. Ore, Deducing event chronology in a cultural heritage documentation system, in : Making History Interactive. Computer Applications and Quantitative Methods in Archaeology, 2010, pp. 122–129. Arcaeopress, Oxford.
[25] J. Searle, Making the social world : The structure of human civilization, Oxford University Press, 2010. doi :0.1093/acprof:osobl/9780195396171.001.0001.

Graphes de connaissances pour représenter et analyser l'évolution des territoires en Histoire

Knowledge graphs to represent and to analyse the evolution of territories in History

Lucas Bourel[1], William Charles[1], Nathalie Hernandez[1,2], Nathalie Aussenac-Gilles[2], Victor Gay[1] and Sébastien Poublanc[3]

[1]*IRIT- CNRS et Université de Toulouse http:// www.irit.fr*

[2]*Université Toulouse2 Jean Jaurès*

[3]*Toulouse School of Economics (TSE) Institute for Advanced Study in Toulouse (IAST) University of Toulouse Capitole*

Abstract

The notion of territory amounts for an important part of human and social sciences. As a spatio-temporal object considered from the digital humanities viewpoint, the question of its digital representation brings forward the need to represent its various aspects. Territory is considered here as a geographical area defined by stakeholders exerting a defined kind of power (religious, legal, ...) over said area, or at least trying to exert it. Inspired by the territorial ontologies TSN and TSN-Change, we propose the HHT ontology (Hierarchical Historical Territory) which is designed to represent the state of affairs of various hierarchical territorial divisions identified by historians' research. Such territory representation implies an overcoming of the main difficulty of such work, which is the lack of exhaustive historical sources regarding the whole territory. Additionally, the HHT ontology endeavors to represent the former states of historical knowledge of territories, now considered deprecated. Defined as part of the ANR ObARDI project, which aims to study territories under the historical period known as Ancien Régime (1661-1789), HHT fills the specific need of representing hierarchical historical territorial divisions.

Keywords

ontology, Semantic Web, digital humanities, spatio-temporal objects

1. Introduction

Dans le domaine des sciences humaines et sociales, la notion de territoire constitue un objet d'étude central à cheval entre l'histoire, la géographie et l'économétrie. Pour ces disciplines, il se caractérise par les éléments suivants :

— un pouvoir exercé par un acteur sur le territoire (**la domination**) ;
— l'espace dominé par ce contrôle territorial (**l'aire**) ;
— la connaissance des **limites** qui enserrent le territoire.

En conséquence, un territoire peut être un découpage administratif, un espace étatique, ou tout espace revendiqué par ses habitants [1]. Il s'agit toujours de l'appropriation d'un espace par un acteur.

Mais un territoire n'est pas réductible à sa seule délimitation spatiale ; la notion intègre également une dimension temporelle car le territoire évolue au cours du temps. Les paroisses de

. *Workshop on Digital Humanities and Semantic Web*

. ✉ lucas.bourel@irit.fr (L. Bourel) ; william.charles@irit.fr (W. Charles) ; nathalie.hernandez@irit.fr (N. Hernandez) ; aussenac@irit.fr (N. Aussenac-Gilles) ; victor.gay@tse-fr.eu (V. Gay)

1. cf http://geoconfluences.ens-lyon.fr/glossaire/territoire

Paris en 1789 sont très différentes des communes - leurs équivalents actuels - tant par leur surface, leurs caractéristiques d'urbanisation, ou leur place dans les nomenclatures administratives. Pourtant, on peut établir un lien de filiation directe entre les paroisses et les communes. Par conséquent, il s'agit de parler d'un objet spatio-temporel dont l'évolution est déterminée par des activités humaines. Enfin, les territoires sont imbriqués dans des rapports de force les uns par rapport aux autres, s'inscrivant dans des hiérarchies territoriales complexes, faisant l'objet de revendications ou de conflits.

Dans une perspective d'humanités numérique, la notion de territoire implique de prendre en compte chacune de ces dimensions afin de pouvoir correctement visualiser et analyser les évolutions territoriales étudiées.

Dans le domaine du web sémantique, les ontologies ont montré leur intérêt pour définir des vocabulaires partagés servant à décrire des entités d'un domaine en vue de les manipuler. Ces descriptions forment des graphes de connaissances. Représenter la notion de territoire à partir d'ontologies présente plusieurs avantages. Tout d'abord, les représentations des territoires réalisées à partir d'ontologies peuvent être facilement interrogées et visualisées. Décrites à partir de vocabulaires partagés, ces représentations de territoires peuvent alors être publiées sur le Web de données liées, partagées et exploitées par les chercheurs de différentes disciplines. Les graphes de connaissances permettent également de représenter plusieurs points de vue d'intérêt différents sur ces territoires, mais aussi de compléter la représentation intrinsèque d'un territoire à l'aide de données connexes (données démographiques, sociales, institutionnelles recensement de la population, présence de rébellions, etc.). Cette démarche s'inscrit dans celle des science ouvertes et collaboratives.

Plusieurs travaux proposent déjà une modélisation des territoires à partir de graphes de connaissances [1, 2, 3, 4]. Dans la plupart de ces approches, on suppose qu'une institution normative (comme l'Union européenne ou l'État français) détermine les territoires en jeu à partir d'une date donnée. Ces territoires sont décrits dans des sources primaires. Par exemple, l'INSEE fournit, sous forme de données ouvertes, la liste des différents types de territoires administratifs français, leur code et leur géographie pour une date donnée [2] [3]. Les territoires sont organisés hiérarchiquement dans la perspective administrative établie par l'institution normative.

Cette approche moderne des modélisations hiérarchiques des territoires ne convient toutefois pas à la représentation d'une organisation hiérarchique territoriale du point de vue historique. En effet, sous l'Ancien Régime par exemple, il n'existe pas de source unique décrivant l'ensemble des territoires et leur organisation hiérarchique pour une date précise et à l'échelle de tout le royaume de France. Les archives utilisées pour reproduire les nomenclatures en usage à cette époque sont forcément partielles, multiples et souvent contradictoires. C'est aux historiens de les interpréter pour arriver à un état des connaissances que l'ontologie se devra de restituer.

A la croisée de l'histoire, de la géographie, de l'économie et des sciences informatiques, l'un des enjeux du projet ANR transdisciplinaire ObARDI est de réussir à modéliser les unités territoriales et leurs hiérarchies entre 1661, début du règne personnel de Louis XIV qui correspond à la mise place de la monarchie "administrative", et 1789, fin de l'Ancien Régime.

2. https://www.insee.fr/fr/information/2016807
3. http://rdf.insee.fr/

Dans le cadre d'ObARDI, l'objectif pour les historiens est d'améliorer la compréhension des dynamiques de pouvoir qui sous-tendent la construction de l'État moderne en France. En étudiant ses mécanismes de développement et sa manière de représenter un territoire politique, il s'agit de dépasser le méta-récit de la construction de l'État qui en freine encore sa compréhension [4] [5] [6]. Pour cela, la représentation à travers une ontologie de ce concept-clé qu'est le territoire est centrale non seulement dans la démarche des historiens mais également dans l'analyse des données territoriales de l'Ancien Régime.

Dans cet article, nous proposons l'ontologie HHT (*Historical Hierarchical Territory*) pour représenter les caractéristiques intrinsèques d'un territoire historique. Cette ontologie permet de représenter plusieurs découpages hiérarchiques simultanés du territoire et l'évolution de ces découpages territoriaux au cours du temps. Elle s'inspire fortement des ontologies TSN et TSN-Change [1] consacrées à la représentation de nomenclatures territoriales organisées de manière hiérarchique, et de leurs changements dans le temps. Nous proposons également une approche itérative pour construire un graphe de connaissances à l'aide de cette ontologie à partir de descriptions de différents unités territoriales, de leurs relations hiérarchiques et de connaissances sur leur évolution selon plusieurs points de vue (administratif, juridique ou religieux par exemple) durant une période historique.

La suite de l'article s'organise en 3 parties. La section 2 présentera la définition choisie des découpages en unités territoriales à travers le temps. La section 3 exposera les fondements de l'ontologie HHT construite à partir de cette réflexion. Enfin, la section 4 décrira la méthode de construction d'un graphe de connaissances à l'aide de cette ontologie.

2. Unité territoriale et hiérarchie

Cette section détaille la définition des unités territoriales et leurs multiples hiérarchies du point de vue de la recherche en histoire.

2.1. Identité d'une unité territoriale à travers le temps

Les historiens attribuent différentes caractéristiques aux territoires, en particulier des liens de filiation, disruptive ou non, que l'on peut définir à travers le temps entre des territoires.

Dans le cadre d'ObARDI, les historiens appréhendent le territoire comme un rapport de force d'un acteur sur un espace géographique. Comme ce rapport de force est amené à évoluer dans le temps, la question se pose alors de l'identité du territoire à travers le temps [7]. Représenter une entité spatio-temporelle inter-connectée à d'autres de ses semblables se modifiant au cours du temps pose de nombreuses questions de modélisation. Plusieurs approches fondamentales existent pour y répondre. Nous reprenons la dualité entre perdurance (propriété des entités qui ne changent pas dans le temps) et endurance (propriété des entités qui ont une durée déterminée, y compris instantanée) proposée par N. Guarino pour structurer les ontologies formelles [8]. La manière d'organiser une ontologie selon ces notions conduit Grenon et Welty à définir deux types d'ontologies [9]. Pour rendre compte de la perdurance, les ontologies SPAN offrent une vision 4D des objets qui sont des "vers d'espace-temps" perdurant dans le temps. Construite pour

4. https://obardi.hypotheses.org/270/

représenter des entités endurantes, les ontologies SNAP donnent une vision tridimensionnelle d'un objet qui dure dans le temps. Il va falloir choisir un camp pour notre objet territorial.

En reprenant les choix retenus dans les travaux sur TSN & TSN-Change[1], l'approche 4D-Fluent (Une approche perdurantiste qui parle de *TimeSlice* pour représenter les versions du perdurant) semble la plus appropriée et correspond à notre définition du territoire établie avec les historiens du projet. Le territoire est vu alors comme un *processus comprenant l'ensemble des tranches de son existence*. La vie du territoire existe à travers toutes les versions de lui-même. Mais chacune de ces versions possède des attributs et des caractéristiques qui lui sont propres. Les historiens s'intéressant particulièrement à la question de l'évolution, le type de changement qui se produit entre deux versions d'un territoire doit être également modélisé dans notre ontologie [5]. Cependant, cette approche soulève deux questions :

— Quelle est le critère d'identité diachronique qui permet d'établir l'identité du territoire ? [7]
— Quelles sont les caractéristiques essentielles d'une unité territoriale qui définissent la singularité de chacune des versions du territoire ?

Nous avons donc cherché à définir les caractéristiques intrinsèques d'un territoire. Les données démographiques, sociales ou les caractéristiques d'urbanisation d'un territoire sont des observations statistiques faites sur le territoire mais ne le caractérisant pas en lui-même. Un échange approfondi avec des historiens ainsi que l'étude des données usuelles associées aux territoires en histoire ont permis d'arriver à d'autres propriétés caractérisant le territoire lui-même :

— Un nom
— Une géométrie (représentant sa délimitation spatiale)
— Son type ou sa catégorie hiérarchique (indiquant son rôle dans une hiérarchie donnée, tel que l'intendance, un diocèse, une élection, etc.)
— Ses relations hiérarchiques.

Le dernier point est particulièrement intéressant car il définit le territoire par rapport à d'autres territoires. Un territoire peut se placer sous la juridiction d'un autre (l'unité territoriale supérieure) et plusieurs autres territoires peuvent tomber sous sa propre juridiction (unités territoriales inférieurs). Ces relations hiérarchiques reflètent encore une fois des rapports de force entre acteurs. Elles sont cruciales car elles vont former le squelette hiérarchique des institutions de l'Ancien Régime, un des principaux objets d'étude dans le cadre d'ObARDI.

Aucune de ces quatre caractéristiques n'est cependant essentielle pour l'identité d'un territoire [10]. Mais, chacune de ces caractéristiques peut engendrer un changement accidentel [7] qui fait passer d'une version du territoire à l'autre. Chacune de ses modifications peut se modéliser sous la forme d'un événement *ChangeBrige* [11] Cela répond à notre deuxième question sur les changements non-disruptifs [7], c'est-à-dire ceux qui ne modifient pas l'identité du perdurant représentant la vie d'un territoire. Reste à déterminer le critère permettant de qualifier un changement disruptif. Toutefois, ce critère ne doit rester qu'un guide. En effet, dans une démarche de recherche historique, seul l'historien peut, grâce à l'analyse des sources, trancher sur la véracité ou non d'une connaissance historique. Néanmoins, il est utile d'expliciter le critère diachronique d'identité que l'on veut justement expérimenter à l'épreuve de l'analyse des

5. L'approche est identique encore une fois à celle de TSN&TSN-Change.

historiens. En reprenant celui développé par Garbazc [7], on énonce le critère d'identité local suivant :

Changement disruptif il y a si et seulement si le nom s'en retrouve modifié en même temps qu'une autre caractéristique du territoire (catégorie, géométrie, relation hiérarchique). Tout autre changement dans lequel serait impliqué un territoire sera non disruptif.

2.2. Relations hiérarchiques

Une hiérarchie territoriale est considérée ici comme une classification des territoires. Elle repose sur un critère hiérarchique identifié par l'historien qui va servir de caractère discriminant permettant d'établir cette classification. Chacune de ces classes possède un niveau hiérarchique, et peut être rattachée à un ou plusieurs niveaux supérieurs et inférieurs. La plupart des découpages administratifs modernes se construisent à l'aide de tel système hiérarchique de découpage du territoire qui sert de nomenclature.

Ces hiérarchies sont le reflet des relations hiérarchiques des territoires qui les constituent. Autrement dit, si une catégorie en domine une autre, une unité de la première catégorie pourra alors dominer une unité de cette deuxième catégorie.

Une différence est à noter entre territoire et unité territoriale. Cette dernière est une catégorie plus générale de la première, définie par les mêmes caractéristiques, mais ne représentant pas historiquement l'appropriation par un acteur de l'espace. Cette distinction est utile lorsque des catégories étiques [6] sont à modéliser.

Par exemple, dans le cadre de la norme européenne NUTS, des seuils démographiques sont fixés par niveau hiérarchique. Pour respecter des réalités socio-économiques, la nomenclature NUTS tente également de respecter les divisions administratives locales déjà mises en place [7]. Cette hiérarchie représente donc à la fois des unités territoriales simples et des territoires.

Le système hiérarchique obtenu découpe alors le territoire européen en plusieurs niveaux :

- Pays membre ;
- NUTS 1 : grandes régions socio-économique. (entre 3 millions et 7 millions d'habitants) ;
- NUTS 2 : régions de bases pour l'application de politiques régionales (entre 800 000 et 3 millions d'habitants) ;
- NUTS 3 : petites régions (entre 150 000 et 800 000 habitants).

L'organisation des unités territoriales sous l'Ancien Régime est plus complexe que les normes européennes actuelles. Après discussion avec les historiens, pour étudier les territoires de l'Ancien Régime, quatre dimensions de classification ont émergé : administrative, religieuse, judiciaire, fiscale. Ces quatre dimensions permettent d'établir quatre découpages hiérarchiques territoriaux différents du royaume de France, quatre filtres d'étude possibles qui se superposent les uns aux autres. Ils ne sont pas étanches entre eux puisque les unités territoriales peuvent posséder chacune une ou plusieurs dimensions.

Chacune de ces dimensions possède ses propres niveaux hiérarchiques. Dans ObARDI, le *découpage administratif* est constitué de pays d'États, de généralités et d'intendances, d'élections et de subdélégations, mais aussi de diocèses civils. Le *découpage judiciaire* est constitué des justices royales que sont les parlements, présidiaux, bailliages et sénéchaussées. Le *découpage*

6. correspondant à un filtre du point de vue de l'observateur et non pas de la réalité historique de l'époque
7. https://ec.europa.eu/eurostat/fr/web/nuts/principles-and-characteristics

religieux contient des archevêchés, des évêché, des archidiaconés, des doyennés et des paroisses ecclésiastiques. Enfin, le *découpage fiscal* contient les greniers à sels des gabelles, les assiettes des tailles, ou encore les départements des tabacs.

2.3. Un état de fait évolutif

Les découpages territoriaux de l'Ancien Régime relèvent d'un état de fait qui atteste de rapports de force entre différents acteurs. Ils ont laissé comme trace de leurs conflits juridictionnels ces hiérarchies territoriales qui structurent l'espace du royaume de France. Analyser les changements que subissent ces unités permet de comprendre la mise en concurrence des pouvoirs sous l'Ancien Régime, les oppositions développées entre acteurs, agents de la puissance publique et les institutions.

Cependant, aucune source ne décrit de manière exhaustive l'état de l'ensemble des territoires étudiés pour une date donnée. Aucune source ne décrit non plus à elle seule l'ensemble des changements subis par ces territoires.

Parallèlement, les historiens souhaitent également documenter l'évolution des connaissances qui a mené à la représentation d'un état de fait à un moment donné. À la lecture de nouvelles sources ou à la réalisation d'enquêtes nouvelles, les représentations changent et se modifient. Il convient donc d'en garder trace dans une perspective méthodologique et épistémologique.

2.4. Etat de l'art

Cette section décrit les ontologies utilisées dans le domaine du web sémantique en lien avec la notion de territoire.

2.4.1. Temps, espace et unités territoriales

L'ontologie OWL-time propose un vocabulaire standardisé par le W3C pour représenter le temps [12]. Elle permet de représenter des instants ou des intervalles temporels dans divers calendriers et d'exprimer des relations topologiques entre eux.

Concernant l'espace, GeoSPARQL propose non seulement un vocabulaire mais aussi un mécanisme de raisonnement spatial [13]. Ce vocabulaire permet de définir à l'aide de coordonnées et d'une forme (un *shapefile*) une zone définie dans l'espace. Néanmoins, dans notre cas, plus qu'une simple zone géographique, nous cherchons à définir un territoire.

Des ontologies d'applications sont spécifiquement dédiées à des découpages territoriaux administratifs. On peut citer ainsi l'INSEE qui propose une ontologie capturant le découpage administratif français actuel pour décrire un jeu de données sur le web des données liées [8].

Néanmoins, ce que nous cherchons est une ontologie de domaine permettant de décrire des découpages territoriaux hiérarchiques génériques. TSN et TSN-Change sont justement deux ontologies construites à cette fin [1]. TSN décrit un découpage pour l'ensemble d'un territoire à une période donnée tandis que TSN-change décrit les changements permettant de passer d'une version de ce découpage territorial à une autre.

8. http://rdf.insee.fr/

Dans TSN, les notions d'unité territoriale, de niveau hiérarchique, et de nomenclature représentant une hiérarchie stable des territoires sont définies.

Contrairement aux agences statistiques auxquelles sont principalement dédiées TSN et TSN-Change, des découpages stables de la France, même pour une courte durée, n'existent pas durant l'Ancien Régime. Chacune des unités territoriales de l'Ancien Régime est documentée par des sources différentes qui ont leur propre temporalité.

2.4.2. Trajectoire de vie

Nous avons évoqué dans la section précédente, la question d'identité d'un territoire à travers ses différentes versions. La *trajectoire de vie* d'une unité territoriale correspond à l'ensemble des différentes versions de cette unité au cours du temps [1]. Dans TSN et TSN-Change, la trajectoire de vie des territoires est rythmée par des versions régulières de la nomenclature toute entière. Dans notre cas, chaque trajectoire de vie peut posséder ses propres références, ses propres sources. C'est sur cette notion différente de la temporalité de ces trajectoires de vie que nous allons devoir diverger de TSN et de TSN-Change.

CIDOC-CRM est une des ontologies de référence pour le patrimoine [14]. Elle a été créée en premier lieu pour répondre à la question d'archivage d'objets culturels, mais elle peut s'étendre hors de ce cadre. Elle est organisée autour de la notion d'évènement. Dans CIDOC-CRM, la notion de zone géographique est définie ainsi que des relations d'inclusion spatiale. Mais, ni les niveaux hiérarchiques ni la notion de territoire ne sont considérés. Cependant les changements territoriaux que nous cherchons à décrire (fusion de deux territoires, apparition d'un nouveau territoire, etc...) peuvent être perçus comme des évènements historiques du point de vue de CIDOC-CRM. Certaines notions que nous cherchons à représenter s'alignent donc avec des concepts de CIDOC-CRM.

Pour représenter les hiérarchies territoriales historiques, nous allons donc devoir créer notre propre ontologie en réutilisant des concepts ou principes de ces vocabulaires.

2.5. Spécification de l'ontologie

Nous présentons ici la démarche suivie pour représenter au sein d'une ontologie les caractéristiques et les spécificités des hiérarchies territoriales historiques que nous venons d'énoncer.

La première étape dans une telle démarche, selon la méthodologie NEON [15] est d'identifier les exigences de l'ontologie. En premier lieu, une discussion permanente avec les experts du domaine, ici les historiens, a permis de cerner plus finement la problématique de cette modélisation par rapport à d'autres. L'objet d'étude et ses caractéristiques ont pu être détaillés dans les sections précédentes suite à ces échanges. Il a fallu également préciser les spécificités des sources disponibles pour les données historiques prévues dans le cadre d'ObARDI [16].

Le *but* de l'ontologie est donc de représenter des hiérarchies territoriales multiples dans le cadre de la recherche historique afin de pouvoir analyser leur évolution spatio-temporelle dans le cas de l'Ancien Régime.

Mais la *portée* de cette ontologie dépasse le seul cas d'étude d'ObARDI pour traiter la question de la représentation d'un territoire en histoire. Le développement d'un portail sémantique

est également prévu afin d'interagir avec cette ontologie, les données du projet et de pouvoir proposer des règles de raisonnements supplémentaires pour ces données.

En terme d'*implémentation*, l'ontologie sera implémentée avec la technologie RDF et le langage OWL.

Les *utilisateurs finaux* de cette ontologie se situeront à la croisée des différentes disciplines impliquées dans le projet ANR ObARDI. Des historiens, des économistes ou des géographes pourront utiliser ce portail afin de manipuler l'ontologie et les connaissances ainsi modélisées.

Les discussions avec les membres du projet ObARDI ont permis d'établir des cas d'usages de cette ontologie ainsi que des *Competency Question* (CQ).

Néanmoins, l'ontologie HHT proposée dans cet article, n'est pas encore à complète maturité. En suivant les méthodes agiles tel que SAMOD [17], nous travaillons par cycles itératifs pour affiner cette ontologie. Voici quelques exemples de CQ relevées pour le moment :

- Comment évoluent les institutions en concurrence au cours du temps ?
- Quelles sont les superpositions de juridictions à un moment donné et à un endroit donné ?
- Quelles sont toutes les paroisses en conflit à une date donnée ?
- Quelle est l'évolution de l'institution X dans la durée ?

L'ontologie HHT nécessite encore plusieurs itérations supplémentaires afin de pouvoir répondre à tous ces cas d'usage. Néanmoins, dans l'état actuel des choses, elle répond à un sous-ensemble des questions soulevées : celles portant sur l'état de fait et sur son évolution.

Dans le suite de la méthodologie NEON, l'ontologie doit s'inscrire dans le réseau grandissant des ontologies déjà existantes à travers différents scénarios. Le scénario 1 [9] a été suivi pour partir des spécifications jusqu'à l'implémentation de l'ontologie HHT. En cherchant des ressources ontologiques supplémentaires, des ontologies décrivant l'espace (GeoSPARQL) et le temps (OWL-Time) ont été intégrées au projet (scénario 3 [10]). HHT s'est concentré alors seulement sur l'aspect territorial et hiérarchique. TSN&TSN-Change a semblé la meilleure piste pour l'aspect hiérarchique, mais des différences fondamentales existant avec le cadre de la recherche historique nous ont poussé à implémenter sous une nouvelle forme une partie de cette ontologie, avec des bases plus générales (Scénario 4 [11]). Concernant les évènements qui retraduisent les changements territoriaux, ces concepts de l'ontologie HHT peuvent étendre la notion d'évènement déjà présent dans l'ontologie de référence en histoire CIDOC-CRM (Scénario 8 [12]).

L'ontologie HHT ainsi produite s'inscrit donc dans le réseau des ontologies déjà existantes sur ce domaine.

Les objectifs auxquels HHT tente de répondre sont alors de modéliser les éléments suivants :

— les unités territoriales ;
— leurs niveau hiérarchiques ;
— différents critères de classification hiérarchique ;
— les changements subis par ces unités et ces niveaux ;
— l'évolution des connaissances des historiens.

9. "From specification to implementation"
10. "Reusing ontological resources"
11. "Reusing and re-engineering ontological resource"
12. "Restructuring ontological resources"

3. Ontologie HHT

Dans cette section, nous détaillons l'ontologie créée pour le projet ObARDI et destinée à représenter des unités territoriales (UT) organisées hiérarchiquement en histoire. Nommée HHT pour *Hierarchical Historical Territory*, elle s'inspire très fortement des ontologies TSN et TSN-Change mais s'en éloigne sur des points fondamentaux.

Cette ontologie repose sur trois modules :

— **Module 1** : La modélisation des territoires et de leurs hiérarchies.
— **Module 2** : Les changements et les modifications appliquées à ces territoires.
— **Module 3** : Les relations de revendication s'opposant à l'état de fait.

Dans la suite de l'article, nous choisirons *hht* comme préfixe de cette ontologie.

3.1. Module 1 : Unités territoriales et hiérarchies

Cette partie de l'ontologie cherche à représenter n'importe quelles hiérarchies territoriales historiques. Elle repose sur différents concepts :

`hht:Area` Simple zone géographique (héritant de `geosparql:Feature` [13] provenant de l'ontologie GeoSPARQL [13]. Ce concept n'est défini que par sa dimension spatiale.

`hht:Unit` sous-classe de `hht:Area`, représentant une zone géographique appartenant à une hiérarchie. Ce concept est défini par sa dimension spatiale et par sa dimension hiérarchique.

Les paroisses : Notons que dans le cadre du projet ObARDI nous ne disposons pas de la géométrie des unités territoriales. Et bien souvent d'ailleurs, en histoire, il est difficile de reconstruire une représentation spatiale précise des lieux décrits dans les sources.

Pour combler ce manque, nous avons identifié un niveau de référence à notre découpage. La géométrie des unités de ce niveau de référence sera considérée comme fixe sur toute la période d'étude. Pour le cas d'ObARDI, ce sont les paroisses qui ont rempli ce rôle [14]. Niveau hiérarchique le plus bas suivant nos quatre critères hiérarchiques, les frontières des paroisses ont très peu changé durant l'Ancien Régime. En considérant la géométrie des unités du niveau hiérarchique le plus bas possible comme fixe, on peut alors reconstruire une notion d'espace discret pour tous ceux qui lui sont supérieurs.

`hht:historicalTerritory` sous-classe d'`hht:Unit`, représentant une portion de l'espace géographique réclamé ou occupé par une personne, un groupe de personnes ou une institution qui en définit elle-même les frontières.

`hht:Level` Niveau hiérarchique permettant de classifier des `hht:Unit`

13. http://www.opengis.net/ont/geosparql{#}Feature
14. https://obardi.hypotheses.org/526

hht:HierarchicalCriterion Critère hiérarchique, caractère discriminant qui permet de définir un découpage hiérarchique de l'espace en différents `hht:Level`. Dans ObARDi, on dispose de quatre critères : administratif, religieux, judiciaire, fiscal.

Chacun de ces concepts permet de décrire l'espace. Pour prendre en compte leur évolution dans le temps, nous définissons 3 nouveaux concepts, versions des concepts précédents, possédant chacun leur propre période de validité (à travers la propriété `hht:validityPeriod`).

hht:UnitVersion Version d'une `hht:Unit` sur une période de validité donnée. Elle possède des unités supérieures auxquelles elle est liée par la propriété `hht:hasSuperUnit` et des unités inférieures auxquelles elle est liée par la propriété `hht:hasSubUnit`.

hht:HistoricalTerritoryVersion Version d'un `hht:HistoricalTerritory` sur une période de validité donnée.

hht:LevelVersion Version d'un niveau hiérarchique sur une période de validité donnée. Un niveau possède un niveau supérieur par la propriété `hht:hasSuperLevel` et un niveau inférieur par la propriété `hht:hasSubLevel`. Un niveau hiérarchique possède des `htt:UnitVersion` à travers la propriété `htt:hasMember`.

La figure 1 représente les concepts ainsi que les relations définies dans le module. Ce schéma est très fortement inspiré de la structure de TSN. Mais, la principale différence réside dans les trajectoires de vie de chaque unité. Dans notre cas, les différentes versions d'une même unité ne dépendent que d'elle-même.

3.2. Module 2 : Changements

Le module 2 vise à représenter les changements ayant mené à la création d'une nouvelle version d'unité territoriale. Trois grands types de changements sont considérés, dont les deux premiers sont inspirés de TSN-Change :

- `hht:FeatureChange` : représentant la modification d'une simple caractéristique entre deux versions d'une unité territoriale : `hht:Expansion`, `hht:Contraction`, `hht:Deformation`, `hht:Disapperance`, `hht:Appereance`, `hht:NameChange`, `hht:UpperUnitChange`, `hht:SubUnitChange`,...
- `hht:CompositeChange` : changement composite, c'est-à-dire un évènement regroupant plusieurs changements simples (`hht:FeatureChange`) : `hht:Merge`, `hht:Split`, `hht:Redistribution`.
  Par exemple, une fusion consiste en la disparition de deux territoires et l'apparition d'un nouveau territoire sur l'espace des deux derniers.
- `hht:UpdateKnowledge` : Cet évènement est différent des deux autres dans le sens où il ne décrit pas un évènement historique mais la mise à jour du graphe de connaissances. Il permet de retracer ainsi l'ensemble des versions par lesquelles est passée l'information sur une unité territoriale pour arriver à l'état actuel des connaissances. Seule la dernière

Figure 1 : Schéma de l'ontologie HHT

est considérée valide tandis que les autres sont considérées comme des connaissances obsolètes.

Chacun de ces concepts est une sous-classe du concept E5 Évènement [15] de CIDOC-CRM .

3.3. Module 3 : Conflits et Revendications

Une troisième partie de cette ontologie concerne la représentation des conflits territoriaux. Analyser l'évolution de ces conflits et les visualiser sont des cas d'usage qui sont revenus souvent de nos discussions avec les historiens du projet.

Dans un premier temps, pour parler de conflits, il faut parler d'acteurs. Comme énoncé précédemment, un territoire est vu comme l'appropriation d'un espace par un acteur. On définit `hht:Institution` comme un sous-type du Concept E39 d'Acteurs [16]. Une institution est l'acteur qui dirige un `hht:historicalTerritory` à travers la relation `hht:attachedTo`.

Chacun des territoires présents dans une hiérarchie peut donc se doter de l'acteur qui représente son institution. Les liens de hiérarchies entre territoires sont alors compris comme des rapports de force entre les acteurs associés. Un territoire placé hiérarchiquement sous un autre tombe sous sa juridiction. Cette relation est à dissocier de celle de propriété. Le territoire n'appartient pas à son territoire supérieur. Ces relations qui sont matérialisées par la hiérarchie

15. https://cidoc-crm.org/Entity/e5-event/version-6.2
16. https://www.cidoc-crm.org/Entity/e39-actor/version-6.2 provenant de CIDOC-CRM

relèvent de ce que les historiens nomment *l'état de fait*. Cet état de fait n'est pas toujours l'image la plus précise de la réalité historique mais traduit l'interprétation par les instances les plus supérieurs de la hiérarchie à une date donnée.

A cet état de fait, on doit donc ajouter la notion de *revendication*. Une revendication provient d'un acteur qui souhaiterait l'établissement d'une nouvelle relation hiérarchique dans l'état de fait. Cet acteur peut être une institution dirigeant déjà un territoire, un groupe de personnes membres d'un territoire ou une personne seule. Dans HHT, cette relation de revendication est réifiée à travers la notion `hht:Claim`. Cette revendication ne dure que durant une période donnée (à travers la relation `hht:validityPeriodOfClaim`) et elle est réalisée par un Acteur à travers la propriété `hht:makeAClaim`. Une revendication cherche à établir une nouvelle relation `hht:hasSubUnit` entre deux territoires. Elle se définit donc par un territoire inférieur (`hht:subTerritory`) et un territoire supérieur (`hht:upperTerritory`) par rapport à la nouvelle relation hiérarchique souhaitée.

Enfin, plusieurs types de `hht:Claim` existent, chacun dépendant de la position de l'acteur source de la revendication :

- `DeclarationUnder` : Une revendication provenant de l'acteur du territoire inférieur qui souhaiterait se placer sous une juridiction plus avantageuse pour lui.
- `ClaimTo` : Une revendication provenant de l'acteur du territoire supérieur qui cherche à placer le territoire inférieur sous sa juridiction.
- `AutonomyRequest` : Une revendication sans territoire supérieur, car cherchant justement la création d'un territoire supérieur pour gouverner le territoire inférieur.

Ces différents types de revendications retenus pour l'instant proviennent d'échanges avec les historiens sur cette question.

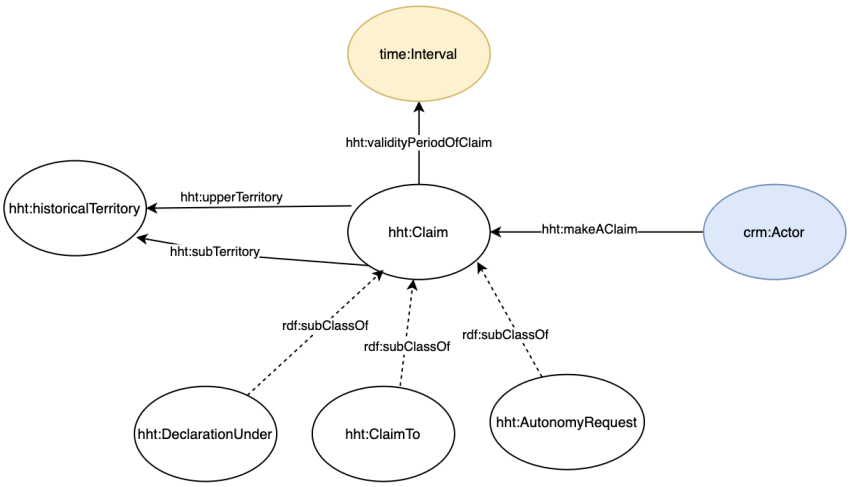


Figure 2 : Schéma du concept de revendication dans HHT

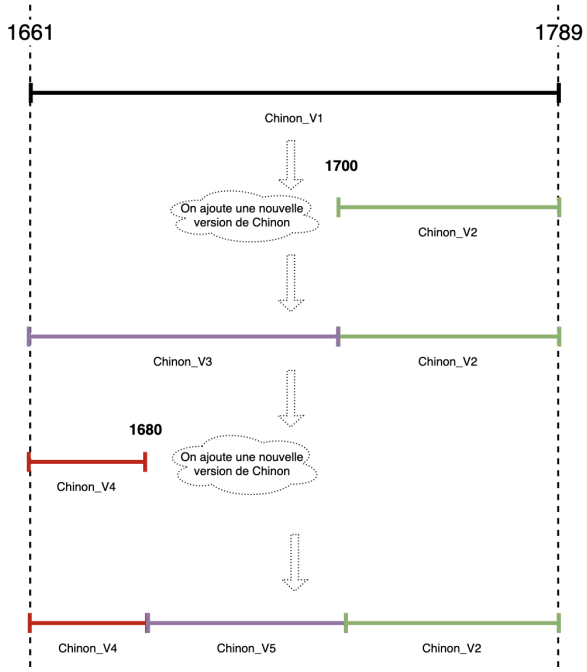


Figure 3 : Différents états pris par la trajectoire de vie d'une unité au fur et à mesure que le graphe de connaissances se peuple.

4. Une construction incrémentale du graphe de connaissances

Le graphe de connaissance d'ObARDI représente l'état de connaissances à propos de l'ensemble des différentes versions des territoires de la France d'Ancien Régime, ainsi que les états de connaissances, désormais obsolètes, par lesquels le graphe est passé. Ce graphe se construit donc de manière continue et incrémentale en lui ajoutant régulièrement de nouvelles ressources.

Au début d'un processus de recherche, on ne fixe que le cadre, et les différents critères hiérarchiques qui seront à l'étude. Ensuite, en consultant les sources, on fixe les différents niveaux hiérarchiques qui correspondent à ces critères. Puis, vient le minutieux et progressif travail de l'analyse des sources pour remplir ce graphe d'unités territoriales.

On considère qu'au premier ajout d'une version d'une de ces unités, la période de référence est par défaut définie sur toute la période d'étude (dans notre cas, 1661-1789) et ceci jusqu'à preuve du contraire.

Lorsqu'une nouvelle version d'unité territoriale vient s'ajouter dans le graphe, les précédentes informations du graphe sont considérées obsolètes. Les différentes versions d'une UT vont donc se modifier en conséquence. Le schéma 3 montre les différents états successifs de la trajectoire de vie d'une unité (ici Chinon) lors de la mise à jour du graphe.

5. Conclusion

L'ontologie HHT permet de représenter des hiérarchies territoriales flexibles et adaptées à la recherche en histoire. Elle ne se concentre que sur l'aspect géométrique et hiérarchique de ces territoires, mais elle permet de les représenter, indépendamment de l'existence d'une source de données décrivant l'entièreté du territoire. Elle facilite une gestion locale des périodes de validité de chacune des unités.

HHT se place dans une perspective historique, s'inscrivant dans les humanités numériques, tandis que TSN est à portée statistique pour proposer une norme commune à tous les découpages territoriaux actuels. Dans cet objectif, HHT cherche également à modéliser les états passés des connaissances historiques. Elle s'aligne aussi avec CIDOC-CRM concernant la notion d'évènement.

Cette ontologie devra être étendue à d'autres concepts pour aider la recherche en histoire. Un aspect non encore modélisé est la notion d'*identité d'un territoire* au-delà des différents critères hiérarchiques. La question se pose en effet de l'existence d'un concept à l'intersection des différentes hiérarchies représentant la même zone.

La notion de revendication *hht:Claim* va également se retrouver centrale lorsqu'il va falloir représenter la notion de front juridictionnel en histoire, ce qui consiste à modéliser des zones géographiques de conflit regroupant plusieurs revendications provenant d'une même cause historique. Modéliser et visualiser ces fronts juridictionnels évoluant au cours du temps est une des finalités du projet ObARDI.

Enfin, un des derniers aspects de la recherche historique dans lequel nous aimerions ensuite poursuivre est la gestion des sources. Le concept d'*HistoricalRessource* dont dépendra la plupart de nos concepts dans HHT, permettra d'indiquer la source de chaque affirmation d'existence d'un territoire, d'un acteur, d'un changement, ou d'un niveau hiérarchique. Dans cette perspective, l'ontologie symogih sera considérée [18].

De nombreuses pistes restent donc encore à explorer pour affiner et étendre ce modèle. L'ontologie sera publiée sur le web ainsi qu'un site dont l'objectif sera de faciliter la construction et la visualisation du graphe de connaissance qui sera partagé par la communauté.

Références

[1] C. Bernard, Immersing evolving geographic divisions in the semantic Web, Ph.D. thesis, Université Grenoble Alpes, 2019.

[2] M. Villanova-Oliver, Représentations de connaissances spatiales évolutives : des ontologies aux géovisualisations, Ph.D. thesis, Communauté Université Grenoble Alpes, 2018.

[3] A. Ezoji, DOTK : territorial ontology as a tool to help the industries for sustainable development JURY, Ph.D. thesis, Université de Technologie de Troyes, 2019.

[4] G. Hiebel, M. Doerr, Ø. Eide, Crmgeo : A spatiotemporal extension of cidoc-crm, International Journal on Digital Libraries 18 (2017) 271–279.

[5] W. Blockmans, The origins of the modern state in Europe : 13th to 18th centuries, Clarendon Press, 1995.

[6] W. P. Blockmans, A. Holenstein, J. Mathieu, Empowering interactions : political cultures and the emergence of the state in Europe, 1300-1900, Ashgate Publishing, Ltd., 2009.
[7] P. Garbacz, B. Szady, A. Ławrynowicz, Identity of historical localities in information systems, Applied Ontology 16 (2021) 55–86.
[8] N. Guarino, Some ontological principles for designing upper level lexical resources, in : First International Conference on Language Resources and Evaluation, Granada, Spain, volume 1, 1998, pp. 527––534.
[9] P. Grenon, B. Smith, Snap and span : Towards dynamic spatial ontology, Spatial Cognition & Computation 4 (2004) 104 – 69.
[10] N. Guarino, C. Welty, Identity and subsumption, in : The Semantics of Relationships, Springer, 2002, pp. 111–126.
[11] T. Kauppinen, E. Hyvönen, Modeling and reasoning about changes in ontology time series, in : Ontologies, Springer, 2007, pp. 319–338.
[12] J. R. Hobbs, F. Pan, Time ontology in owl, W3C working draft 27 (2006) 133.
[13] R. Battle, D. Kolas, Geosparql : enabling a geospatial semantic web, Semantic Web Journal 3 (2011) 355–370.
[14] G. Bruseker, N. Carboni, A. Guillem, Cultural heritage data management : the role of formal ontology and cidoc crm, Heritage and Archaeology in the Digital Age (2017) 93–131.
[15] M. C. Suárez-Figueroa, A. Gómez-Pérez, M. Fernández-López, The neon methodology for ontology engineering, in : Ontology engineering in a networked world, Springer, 2012, pp. 9–34.
[16] V. Gay, S. Poublanc, J.-L. Demonsant, Plan de gestion de données du projet ANR ObARDI, Ph.D. thesis, MSHS Toulouse ; FRAMESPA ; TSE-Toulouse School of Economics ; IAST-Institute ..., 2021.
[17] S. Peroni, A simplified agile methodology for ontology development, in : OWL : Experiences and Directions–Reasoner Evaluation, Springer, 2016, pp. 55–69.
[18] F. Beretta, L'interopérabilité des données historiques et la question du modèle : l'ontologie du projet SyMoGIH, Presses universitaires de Paris Nanterre, 2017. URL : https://halshs.archives-ouvertes.fr/halshs-01559816.

Sur les épaules d'un géant : utilisation des corpus et outils numériques pour l'histoire des discours antimodernes sur l'Europe dans la presse suisse 1900-1945

On a giant's shoulders: using corpus and digital tools for the history of anti-modern discourses on Europe in the Swiss press 1900-1945

Estelle Bunout[1]

[1]*Luxembourg Centre for Contemporary and Digital History (C2DH), Université du Luxembourg*

Abstract

Ce cas d'étude propose de montrer comment l'environnement de recherche créé par la numérisation des sources, la popularisation d'outils de traitement automatisé des textes et surtout le dialogue interdisciplinaire, font évoluer la pratique de la recherche en histoire. Au-delà de l'automatisation de certaines tâches typiques de la recherche historique, comme l'aide à la constitution d'un corpus de recherche ou l'annotation de ce corpus, la réflexion même sur les sources et leur contenu est stimulée par cet environnement. En l'occurrence : une analyse des discours se traduit par la mesurer une présence, ou l'estimation de la représentativité d'éléments distinctifs d'un discours. Ces tâches se prêtent particulièrement à une automatisation partielle, par le biais d'analyse automatisée des textes, rendue possible par leur numérisation. Ce papier présente ées étapes d'opérationalisation d'une analyse de discours en histoire, utilisant un corpus de presse numérisée, des outils de traitement des langues naturelles pour collecter un corpus de recherche, le classer et l'organiser par degré de similarité. L'itération entre conceptualisation, opérationalisation et analyse des résultats ouvre un nouvel angle d'observation des sources historiques.

Keywords

history, digital press, naive Bayesian classifier

1. Introduction

Avant de présenter nos étapes d'automatisation partielle de l'analyse de discours, il nous faudra dans un premier temps définir comment le discours antimoderne se matérialise dans la presse et trouver des textes « similaires ». L'analyse de discours implique un changement d'échelle entre la production d'idées et une certaine représentativité ou circulation de ces idées. Dans le même temps, ce changement d'échelle et la médiation qu'il implique risque également de décupler nos biais et de ne valider que les hypothèses émises au départ, en d'autres termes, ne va-t-on trouver que ce qu'on cherche ? La presse numérisée constitue pour cela, une source potentiellement formidable pour procéder à ce changement d'échelle, permettant de chercher à la fois de manière ciblée et de collecter une grande quantité de texte, simplement par un mot-clé.

Workshop on Digital Humanities and Semantic Web
estellebunout@gmail.com (E. Bunout)
https://www.c2dh.uni.lu/people/estelle-bunout (E. Bunout)

Il faut toutefois trouver des stratégies pour analyser cet accès facilité : comment identifier un discours dans la masse d'articles, publicités et annonces publiés au quotidien ?

C'est la richesse de cette source, que de présenter un kaléidoscope de l'usage du terme « Europe » et c'est un avantage indéniable, de sa forme numérisée, que de pouvoir mesurer relativement simplement, la proportion de ses différents usages et leur évolution au fil du temps. Le passage à une analyse de discours nécessite un traitement particulier de la masse collectée. Nous allons présenter comment la médiation par les outils de fouille de texte automatisée, implique une certaine transparence dans la démarche heuristique, dimension mise en avant dans la recherche en humanités numériques. L'aspect que nous voudrions souligner ici, est l'itération entre la formulation d'hypothèses de recherche, leur transposition en consignes aux algorithmes utilisés et la redéfinition des concepts utilisés [1]. En d'autres termes, comment la confrontation médiée ou distante à la source historique numérisée permet de poser des questions à des sources massives en offrant une interaction adaptée à leur format et envergure.

2. Annoter le corpus « Europe » dans la presse suisse : diversité thématique et discursive

Les articles qui ont servi dans cette analyse, ont été collectés via l'application impresso[1], en utilisant comme mot-clé « europ* », de façon à collecter les articles contenant tant les mots « Europe », qu' « européen », « européenne » etc., pour une diversité de titres de la presse francophone suisse (*l'Express, Gazette de Lausanne, l'Impartial, Journal de Genève, le Confédéré, l'Essor, la Liberté, la Lutte Syndicale, la Sentinelle, Solidarité*). Parmi ces titres, on retrouve des feuilles d'avis, dont la fonction était originellement la diffusion d'informations commerciales, des organes de partis ou syndicats et enfin, des titres de presse d'opinion. Au total, 227 351 articles ont été ainsi collectés.

Pour avoir une première impression du contenu des articles collectés, nous avons eu recours au topic modelling, qui est une méthode de calcul de probabilité de cooccurrence des mots dans une collection de document, sur la base d'observation partielle. Les informations produites par cet algorithme sont d'une part, les mots qui cooccurrent (probablement) au sein d'un document, qui forment un *topic* et la distribution de la présence des *topics* au sein de la même collection de documents. Cet outil permet d'avoir une description du contenu d'une vaste collection de documents, sans prédéfinir les thèmes mais en partant du contenu des documents. Par ces propriété et sa relative simplicité d'usage, il est particulièrement populaire pour la recherche dans la presse numérisée[2]

La détection de cooccurrence de mots au sein d'un document est moins simple à interpréter que les collocations, où on mesure une cooccurrence au sein d'une même phrase, mais jette un filet plus large de l'association récurrente de termes. On peut ainsi utiliser cet outil non

[1] Application développée dans le cadre du projet *impresso*: Media Monitoring of the Past, une colloboration entre les Universités de Zurich et du Luxembourg, et l'Ecole Polytechnique Fédérale de Lausanne, regroupant des collections de presse numérisée suisses et luxembourgeoise, enrichies par du *topic modelling*, la reconnaissance d'entitées nommées, la détection de textes dupliqués (*text reuse*), le plongement de mots (*word embedding*) et présentés dans un interface de recherche et d'exploration commun, accessible sous : https://impresso-project.ch/app/

[2] Comme l'indiquait déjà Blei [2] et comme est discuté dans Bunout et al. [3] à paraître.

seulement pour détecter des « thèmes » mais aussi des styles ou ton d'un texte.

Pour illustrer ce principe, prenons un exemple : le *topic* numéro 40, calculé pour le journal la Liberté se présente ainsi : « nos vos patrie sommes chers applaudissements devons avons avez voulons devoir noire dieu peuple-suisse confiance salut êtes messieurs bravos nous-mêmes ».

Ce sont les mots les plus récurrents pour ce *topic*. Ils forment une association qui reste à interpréter par le chercheur. Ici, on peut supposer qu'il s'agisse de discours prononcés lors de célébrations patriotiques en Suisse. On peut maintenant regarder les articles qui ont reçu une probabilité de contenir ce *topic* et observer parmi eux, un article du 15.09.1934 sur le Jeûne fédéral, des discours reproduits dans la presse, prononcés à l'occasion du Nouvel an de 1942 ou encore des articles de la rubrique d'annonces de contenus de la revue Semaine catholique. Ce qui n'était pas visible dans les mots décrivant le *topic* était donc la dimension religieuse des textes regroupés par celui-ci. Il est important de vérifier « manuellement/lecture humaine » le contenu des sous-collections d'articles créées par cet outil [4].

Les *topics* ont été utilisés pour cibler et sélectionner des exemples archétypaux, pour alimenter l'outil suivant : le classificateur naïf Bayésien. Le principe de ce classificateur est assez simple : sur la base d'une série de texte exemples et de textes contre-exemples, l'algorithme détermine une liste de mots qui est la plus discriminante pour un identifier l'une des deux catégories. Sur la base de sa présence dans l'une ou l'autre ou dans les deux catégories, chaque mot se voit attribuer une valeur de prédiction qui est utilisée pour mesurer la probabilité de textes inconnus (non utilisés pour définir ces catégories) à appartenir à l'une ou l'autre des deux catégories. Pour préparer le corpus d'entrainement, il est conseillé de choisir une diversité relative, pour couvrir les différents aspects d'une même catégorie, dont nous cherchons à identifier d'autres matérialisations. De manière symétrique, le corpus « neutre » ou de contraste, doit lui couvrir une grande diversité d'éléments à exclure ou minimiser dans le classement. Spontanément, on pourrait penser que choisir à partir d'un groupe d'article partageant le même *topic* pourrait donc s'avérer problématique, si la sélection reste trop homogène et ne permet pas de découvrir d'autres articles. Pour mieux comprendre comment cette question de l'« homogénéité » ou « diversité » a été traitée, nous allons détailler les critères retenus pour chaque catégorie et redonner quelques exemples pour illustrer les choix. L'établissement et la redéfinition des catégories de recherche résulte, sans trop d'originalité, d'un retour entre étude des sources et de la littérature scientifique. L'espoir ici est de tenter de matérialiser une pratique analogue : trouver des textes similaires au texte initialement repéré, estimer la proportion de la présence de textes similaires pour juger de l'importance du phénomène que le texte initial retenu reflète. Face à la masse de texte et l'impossibilité de redonner l'ensemble des sources pertinentes au lecteur, les analyses de discours se trouvent souvent amoindries au moment de la restitution des résultats.

Nous voudrions ici faire un chemin inverse à cette restitution traditionnelle en partant des idéaux-types que nous avons sélectionnés et discuter de la définition et mesure de la « similarité » via un classificateur naïf Bayésien (NBC)[3]. Ce faisant, nous rendons explicite les éléments utilisés pour définir l'idéal-type de chaque catégorie. Pour donner ici un exemple de cette démarche, nous nous concentrons sur la catégorie « diplomatique », la moins ambivalente. Pour cette catégorie, nous avions décidé de retenir les articles rapportant des faits avec des

[3]Pour une présentation plus détaillée des étapes, voir [5]

commentaires minimaux sur des rencontres diplomatiques, politiques à l'échelle européenne, ou encore faisant la chronique sans valorisation particulière, des initiatives diplomatiques de coopération européenne, notamment celle d'A. Briand en 1929. Le matériel collecté pour définir cette catégorie se compse d'éditoriaux, notamment de Maurice Muret, parus dans sous le titre « Bulletin politique », ou d'articles paraissant dans des rubriques type « vie internationale ». Ces rubriques sont souvent constituées de brèves et dans la reconnaissance automatique des contours des articles, produite lors de la numérisation de la source, ce type de format est souvent identifié comme un article entier. On pourrait considérer que ce type d'article contient beaucoup de « bruit » et pourrait fausser nos mesures de similarité. Nous avons cependant choisi de garder la rubrique entière pour « entrainer » la classification, car la collection dans laquelle nous allons chercher des articles similaires, en sera également composée. Il y aura donc des articles qui ne traitent pas d'« Europe » mais restent dans la tonalité que nous cherchons à identifier. Nous avons également sélectionné des articles qui rendent compte de conférences d'organisations telles que la Fédération des Unions intellectuelles portant des titres du type « La civilisation européenne en danger », qui aurait pu indiquer une thématique antimoderne, mais sont contenues dans des articles sans commentaires, plutôt informatifs. Enfin, nous avons ajouté des articles contenant des commentaires de type « géopolitique », sur les ambitions hégémoniques d'un pays particulier par exemple. Ainsi, les premiers mots de la liste des mots qui caractérisent cette catégorie semblent refléter le contenu ciblé : « convention », « union-douanière », « accords », « souscommission », « locarno », « délégué » etc. Le mot « convention » est accompagné d'une probabilité 95%, en d'autres termes, sur la base des articles donnés pour l'entrainement, un article qui contient ce mot a une très forte probabilité d'appartenir à la catégorie « diplomatie », et ainsi de suite pour tous les mots identifiés dans les articles du corpus d'entrainement.

La préparation de ces quatre catégories a donc été l'occasion de matérialiser des définitions et de rendre cette matérialisation communicable. Ces collections d'articles servent de base pour mesurer une similarité textuelle avec l'ensemble des articles contenant simplement le mot *europ**. Nous allons maintenant nous pencher sur les résultats de cette mesure de similarité.

3. La similarité lexicale comme brique de l'analyse discursive

Qu'en est-il des résultats de cette mesure de similarité ? La première information à retenir est la proportion d'articles annotés par cette mesure. Pour une première vérification manuelle, nous avons retenu les articles ayant reçu une mesure de 100%, c'est-à-dire que la probabilité d'appartenir à la catégorie mesurée est de 100%. Chaque article étant soumis séparément à la mesure de chaque catégorie (diplomatique, fédéraliste, utopique, antimoderne), il peut recevoir potentiellement une probabilité d'appartenir à chaque catégorie simultanément. Nous reviendrons sur cette dimension plus tard, mais dans un premier temps, il nous faut souligner qu'une partie réduite des articles ont été retenus pour être manuellement inspectés.

Dans la figure 1 ci-dessus, nous voyons que la proportion d'articles annotés est la plus forte en 1922, et la moins pour la période précédente. La quantité absolue reste en revanche la plus importante pour les années 1930 et 1940. Ceci signifie que les textes choisis pour incarner les idéaux-types ressemblent moins aux textes des années 1900-1920. Cette défection de l'utilisation

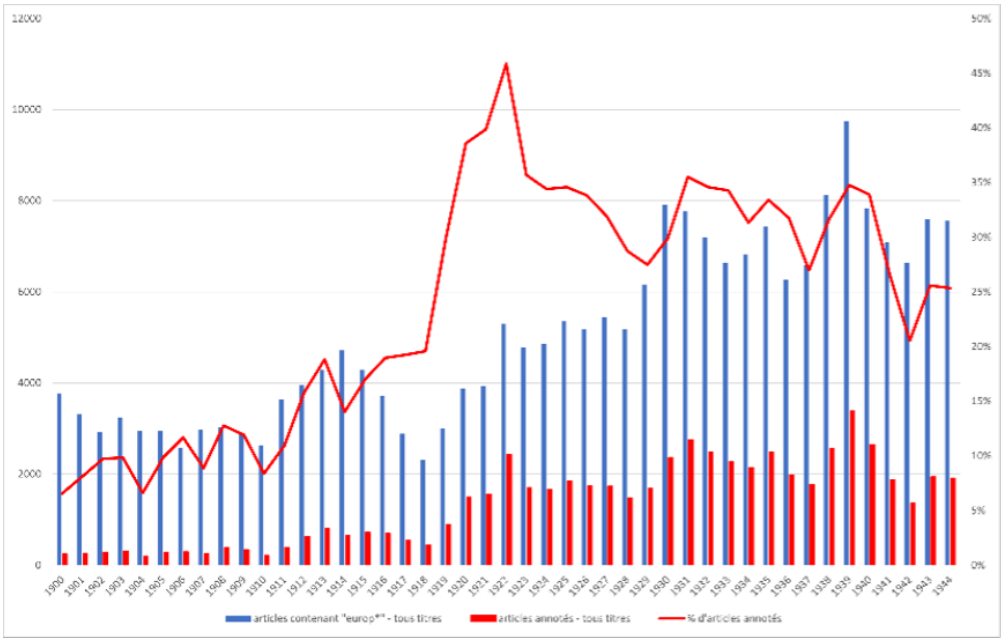


Figure 1: Distribution chronologique en valeurs absolue et relative des articles annotés, pour tous les titres de presse sélectionnés. La colonne bleue représente la totalité des articles collectés, contenant le mot-clé « europ* », la colone rouge, la totalité des articles annotés par au moins un classificateur, tandis que la courbe rouge indique la proportion des articles annotés dans les articles collectés, le tout par année. Par exemple, pour l'année 1944, un peu moins de 8000 articles ont été collectés, un peu moins de 2000 articles ont été annotés, correspondant à environ 25% des articles collectés.

de l'outil peut tout de même nous indiquer une évolution de vocabulaire à ce moment et appeler à une itération supplémentaire reflétant ces spécificités chronologiques, en utilisant des articles sélectionnés dans la même décennie où la similarité sera recherchée. Nous avons opté pour une recherche de similarité tout au long de la période, pour justement forger un point de repère commun pour ces cinq décennies. Toujours dans cette optique de vue d'ensemble, regardons à présent la distribution par titre des annotations.

Dans la figure 2, nous pouvons voir les degrés de similarité des articles retenus pour définir chaque classificateur avec les articles collectés pour chaque titre, contenant le terme « europ* », ce que nous voudrions utiliser pour déterminer la présence de discours respectivement utopique, antimoderne, diplomatique ou fédéraliste sur l'Europe. On remarque tout de suite que le point de référence commun produit une image très différenciée de la similarité pour chaque titre. On ne peut pas en conclure une domination d'un discours antimoderne pour l'Essor ou diplomatique pour le Journal de Genève, mais on peut considérer ce résultat comme indicatif d'une différence de ton, de vocabulaire utilisé dans ces différents titres (tout en gardant à l'esprit que ces articles ne sont qu'une partie des articles parus à ce moment dans ces titres). On voit aussi que les titres qui sont plus représentés dans la collection d'articles idéaux-types, ont une plus grande proportion d'articles annotés à 100%. Notamment, la Sentinelle score remarquablement faiblement, ce qui appelle à une itération supplémentaire nécessaire, reflétant plus nettement ce titre dans les

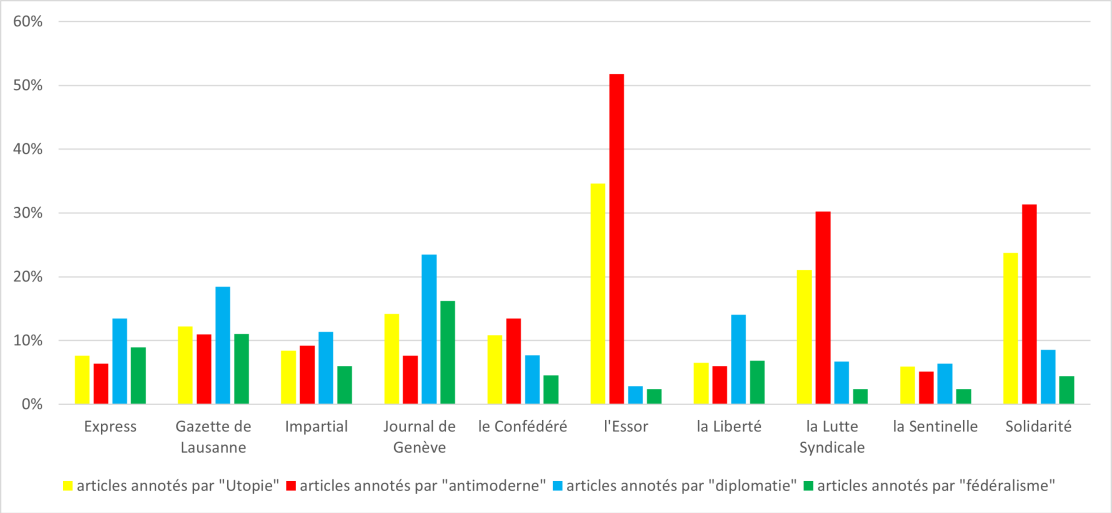


Figure 2: Proportion des annotations d'articles par les quatre classificateurs (utopique, antimoderne, diplomatique, fédéralisme) dans la totalité des articles collectés, par titre.

idéaux-types, et en particulier ses contenus d'utopie socialiste. D'autres titres, tels Solidarité, qui sont également faiblement présents dans la collection initiale, sont réceptifs à la mesure de similarité (ou du moins dans des proportions comparables ou supérieures aux titres mieux représentés, tels la Gazette de Lausanne). On remarque aussi nettement une présence forte de l'annotation « utopique » lorsque l'annotation « antimoderne » est forte, conformément à nos attentes concernant le ton, un peu plus surprenante du point de vue du contenu. Il semble que le classificateur soit plus sensible au style ou ton qu'au contenu. Il nous faut maintenant revenir à la distribution dans le temps de ces annotations, d'abord tous titres confondus, ensuite, nous avons sélectionnés quelques titres présentant des distributions distinctives.

L'espoir de ce travail de collecte et de mesure de similarité était que l'utilisation de ces exemples divers, dont serait extrait automatiquement le vocabulaire saillant, pourrait aider à identifier des articles contenant des discours similaires. Pour nous aider à traiter la masse qui, malgré sa réduction drastique, reste importante (de 227 351 à 60 348 articles), nous avons de nouveau recours aux annotations du topic modelling, présenté plus tôt. Le recroisement des quatre annotations du classificateur et du topic modelling permet de constituer des petits groupes d'articles, au sein de chaque titre, dont le contenu partage a priori certaines caractéristiques de thème et de style. L'accumulation de ces annotations sont utilisées comme faisceaux d'indice pour guider la lecture et l'analyse, avec toutes les précautions mentionnées au cours de cette présentation.

Ainsi, pour le Confédéré, le *topic* 32 est décrit par des mots « européenne souveraineté droit etats briand l'union fédération droit-international l'union-européenne… » et regroupe 14 articles contenant au minimum 30% de ce *topic*. Il aurait déjà retenu notre attention en tant que tel, mais les annotations invitent à interroger la diversité interne de ce groupe d'article : avec 5 articles annotés de « fédéralisme », « antimoderne » et « utopique », tandis que 8 sont annotés comme « diplomatique », « fédéralisme » et « utopique ».

Si les groupes d'articles du Confédéré restent de taille raisonnable, ceux de l'Express sont plus volumineux et les indices apportés par les classificateurs, plus utiles dans ce contexte. On peut ici utiliser cette accumulation d'annotations pour sélectionner par exemple, les *topics* qui contiennent la plus forte proportion d'annotation « antimoderne » par exemple, ou sélectionner parmi les autres qui paraissent pertinent, ceux qui ont cette même annotation. Par exemple, le *topic* décrit par « prix domicile suisse abonnements » n'a reçu aucune annotation d'aucun classificateur. Sa description avait déjà permis de potentiellement exclure de la vérification les 1706 articles qu'il regroupe, mais cette absence d'annotation permet de confirmer la probabilité que ces articles réfèrent à l'Europe comme simple espace de tarification pour des abonnements. Il en va de même pour les *topics* « bourse banque neuchât crédit suisse » et « concert musique disques », regroupant respectivement 1357 et 1280 articles. Par contraste, le *topic* « gouvernement une londres aux France », rassemblant 1271 articles, qui pouvait également sembler couvrir des thématiques diverses, ne contient aucun article annoté à 100% comme « antimoderne » mais 582 articles annotés comme « diplomatiques » et « fédéralistes », qui pourraient être vérifiés. Plus utile encore, le *topic* décrit par ces mots « faire politique lui contre pays guerre » et comptant 2300 articles, compte 389 articles annotés comme « antimodernes ».

Petit à petit, par cette sélection guidée par les annotations et les hypothèses que leur préparation a soulevé, s'ajoutant aux questions initiales, nous pouvons accumuler ces groupes d'articles, qui peuvent rassembler des articles venant d'une rubrique récurrente ou d'une chronique qui est ainsi identifiée, et dont le groupe initial pourra être élargi, ou autour d'événements, comme une discussion parlementaire, ou des rituels politiques, tel le tir fédéral. Certains groupes d'articles ne s'avèrent pas être caractérisés une homogénéité reconnaissable à l'inspection et sont écartés.

Cette démarche aide cependant à dépasser la citation anecdotique et mesurer plus clairement, de manière plus transparente, comment un discours se propage dans un corpus de textes d'archives, comment il cohabite, se superpose à d'autres discours et au contraire, comment il se distingue d'autres. En l'espèce, de la fabrication des catégories et des premières analyses de ces résultats naissent la nécessité de construire une autre mesure de similarité avec des sous-catégories jusqu'à présent intégrées aux quatre catégories étudiées jusqu'alors. Il semble aussi plus facile de distinguer les articles rapportant les événements diplomatiques ou liés aux efforts de coopération institutionnelle européenne des discours à visée utopique ou contenu antimoderne, que de distinguer ces deux dernières catégories. Et finalement, il apparaît que des textes au contenu antimodernes, mentionnant l'Europe au détour d'une phrase, sont plus fréquents que des discours visant à promouvoir une conception antimoderne de l'Europe.

Remerciements

Ce travail a pu être mené grâce aux efforts de numérisation des bibliothèques nationales au Luxembourg et en Suisse, au projet *impresso*[4], aux plateformes de popularisation des outils de traitement de textes, et finalement un script né de longs échanges avec Milan van Lange, chercheur au NIOD (NL)[5].

[4]https://impresso-project.ch/app/
[5]https://www.niod.nl/en/staff/milan-van-lange

Références

[1] D. Nguyen, M. Liakata, S. DeDeo, J. Eisenstein, D. Mimno, R. Tromble, J. Winters, How we do things with words: Analyzing text as social and cultural data, Frontiers in Artificial Intelligence 3 (2020). URL: https://www.frontiersin.org/articles/10.3389/frai.2020.00062/full#h3. doi:10.3389/frai.2020.00062.

[2] D. M. Blei, Probabilistic topic models, Communications of the ACM 55 (2012) 77–84. URL: https://dl.acm.org/doi/10.1145/2133806.2133826. doi:10.1145/2133806.2133826.

[3] E. Bunout, M. Ehrmann, F. Clavert (Eds.), Digitised Newspapers – A New Eldorado for Historians?: Tools, Methodology, Epistemology, and the Changing Practices of Writing History in the Context of Historical Newspapers Mass Digitization, De Gruyter Oldenbourg, 2022. URL: http://www.degruyter.com/document/isbn/9783110729214/html?pds=422022164415299843667093853399 6, publication Title: Digitised Newspapers – A New Eldorado for Historians?

[4] J. Grimmer, B. M. Stewart, Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts, Political Analysis 21 (2013) 267–297. URL: https://www.cambridge.org/core/journals/political-analysis/article/text-as-data-the-promise-and-pitfalls-of-automatic-content-analysis-methods-for-political-texts/F7AAC8B2909441603FEB25C156448F20. doi:10.1093/pan/mps028.

[5] E. Bunout, Grasping the anti-modern discourse on europe in the swiss digitised press, or can text mining generate a research corpus from an article collection?, Journal of Open Humanities Data 7 (2021) 21. URL: http://openhumanitiesdata.metajnl.com/articles/10.5334/johd.37/. doi:10.5334/johd.37, number: 0 Publisher: Ubiquity Press.

Graphes de connaissances pour les humanités numériques : besoins spécifiques et problèmes généraux

Knowledge graphs for the digital humanities: specific requirements and general issues

Mathieu d'Aquin[1]

[1]*Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France*

Abstract

A lot has been written about the benefits and difficulties associated with the creation and use of knowledge graphs and of the broader Semantic Web technologies in the context of the digital humanities. Concrete feedback from projects using such technologies is however rare and, by nature, often focuses on the specific use of knowledge graphs in a particular context. In this article, we describe the use of Semantic Web technologies in three different domains, from three (multidisciplinary) projects. The goal is to better understand how, despite the variety of needs and requirements from researchers in those projects, shared benefits and issues appear. Identifying those shared elements can be helpful in the sense that it can guide the development of Semantic Web tools towards greater relevance and efficacy for the digital humanities.

Keywords

knowledge graphs, Semantic Web, music history, iconography, literature

1. Introduction

Le Web sémantique peut être vu comme une utilisation des technologies du Web pour rendre les données et les connaissances plus accessibles et manipulables au travers de graphes de connaissances (voir [1]) inter-connectées et navigables par des requêtes HTTP. S'ajoute à cela la spécification de la signification de ces données et de ces connaissances par l'utilisation d'ontologies (voir par exemple [2]). Ce type de représentation permet d'encoder les informations relatives à un domaine non seulement de façon à être connecté avec d'autres sources d'information, mais aussi au travers de représentations flexibles et évolutives.

Du fait de leur nature pluridisciplinaire et des sujets complexes qui y sont étudiés, les informations manipulées par les disciplines des humanités numériques sont souvent très riches et variables. Elles incluent un grand nombre de dimensions, instanciées de façon plus ou moins complète, et peuvent faire référence à des sources externes, telles que des bases de données de référence dans le domaine. Pour cette raison, les humanités numériques ont été depuis longtemps considérées comme un champ d'application privilégié pour les technologies du Web sémantique (voir par exemple [3]).

Beaucoup a été dit sur les bénéfices et les difficultés rencontrés dans la mise en place de ces technologies dans des applications des humanités numériques. De nombreux cas d'utilisation

. *Workshop on Digital Humanities and Semantic Web*
. mathieu.daquin@loria.fr (M. d'Aquin)
. https://mdaquin.github.io/ (M. d'Aquin)

ont été décrits, mais les retours d'expérience concrets restent rares [1]. De plus, à part quelques exceptions où les sujets généraux de l'utilisation des technologies du Web sémantique ont été explicitement étudiés comme dans ([4]), ces retours par nature ont tendance à se focaliser sur une utilisation spécifique des graphes de connaissances dans un domaine particulier.

Cet article décrit l'utilisation des technologies du Web Sémantique dans trois domaines différents, issus de trois projets pluridisciplinaires différents :

L'iconographie avec un projet de portail fondé sur un graphe de connaissances décrivant les fresques murales d'églises de Crète.

L'histoire de la musique avec un projet d'acquisition d'une base de milliers de descriptions d'expériences d'écoute de la musique.

La littérature avec l'analyse sémantique des écris de certaines des premières femmes philosophes.

L'objectif est de comprendre comment, malgré des besoins et des attentes variés des chercheurs de ces domaines, émergent aussi bien des avantages que des problèmes communs. Le but est donc de discuter, d'une façon indépendante du domaine d'application spécifique, quels sont les avantages réels tirés des technologies du Web sémantique, et comment celles-ci doivent encore évoluer pour devenir de meilleurs outils pour les humanités numériques.

2. Graphes de connaissances et humanités numériques : avantages attendus et blocages

De façon générale, l'idée des graphes de connaissances (voir [1]) est de représenter les données, informations et connaissances sous la forme de graphes orientés et étiquetés. Les graphes de connaissances suivent la vision générale du Web sémantique en s'appuyant sur ces technologies (HTTP, RDF [2]) pour représenter ces graphes de façon à les rendre compatibles avec une diffusion sur le Web, et avec la possibilité de les interconnecter globalement avec d'autres graphes. Ceux-ci sont aussi qualifiés de graphes de connaissances du fait que la signification des éléments d'information inclus peut être rendue explicite et interrogeable au travers d'ontologies (voir [5]), permettant ainsi de faciliter l'échange de connaissances et le raisonnement sur ces connaissances.

Comme décrit plus haut, de par leur flexibilité et leur ouverture, les approches fondées sur les graphes de connaissances ont gagné en popularité dans plusieurs domaines des humanités numériques depuis de nombreuses années. Dans ces domaines (voir par exemple [3]) et dans des domaines connexes (voir par exemple [6]), les avantages attendus souvent cités inclus :

Applications intelligentes : l'idée de représenter les connaissances au travers d'ontologies provient du domaine des systèmes à base de connaissances, c'est-à-dire d'un certain paradigme de l'intelligence artificielle s'appuyant sur la représentation explicite de connaissances et sur le raisonnement artificiel pour faciliter la prise de décision. Un des avantages souvent cité du Web sémantique est donc de permettre la construction, en suivant ce

1. Voir par exemple le workshop WHISE http://whise.cc/
2. https://www.w3.org/TR/rdf11-concepts/

paradigme, d'applications intelligentes qui exploitent les connaissances incluses dans les graphes de connaissances construits et dans ceux auxquels on se connecte (voir par exemple [7]).

Accès à l'information : à un niveau plus bas, le simple fait de représenter les informations et les connaissances du domaine d'une façon compatible avec les technologies du Web apparaît comme un avantage. Ces technologies sont ouvertes et conçues pour faciliter l'accès de n'importe en ligne, sans contrainte logicielle, ce qui les rend attractives pour des activités de recherche utilisant ces connaissances en comparaison avec l'utilisation de systèmes plus fermés tels que des bases de données relationnelles.

Données liées et intéroperabilité : une des notions centrales du Web sémantique est de permettre de connecter ses propres données avec d'autres, de la même façon que l'on connecte des pages Web entre elles. Cela semble particulièrement utile dans le contexte des humanités numériques considérant que des données de référence existent sous forme de graphes de connaissances (voir par exemple [8]) et peuvent donc être réutilisées. Ces graphes de connaissances utilisant non seulement des technologies standards, mais aussi des vocabulaires et ontologies partagées, il devient possible de créer de nouveaux graphes de connaissances syntaxiquement et sémantiquement interopérables : ils peuvent être utilisés conjointement dans des applications sans difficultés majeures.

Visibilité : beaucoup de projets d'humanités numériques utilisant les graphes de connaissances sont soit liés au patrimoine culturel, soit liés à des activités de recherche dans le domaine concerné. Dans les deux cas, en plus de l'avantage de pouvoir se connecter à des sources d'informations externes comme décrit ci-dessus, l'utilisation des technologies du Web sémantique a aussi généralement pour but de permettre la création de nouvelles ressources de référence, réutilisables par d'autres, et permettant ainsi une plus grande visibilité pour le projet.

Malgré ces avantages attendus, l'utilisation des technologies liées aux graphes de connaissances reste limitée au sein des humanités numériques, et cette utilisation n'est pas toujours soutenue ou pérenne. Dans sa thèse soutenue récemment, [4] s'est intéressée aux raisons pour lesquelles les chercheurs de certains domaines des humanités numériques utilisent ou n'utilisent pas les technologies du Web sémantique et quelles sont les directions possibles pour les améliorer. Parmi les blocages identifiés sont inclus :

Des technologies compliquées et mal connues : un des désavantages les plus communément cités des graphes de connaissances est qu'ils font appel à des technologies qui restent difficiles à comprendre. En effet, celles-ci sont généralement plus récentes que les alternatives possibles, moins bien documentées, et utilisables au travers d'outils souvent développés au sein d'équipes de recherche ne disposant pas de service d'aide suffisant pour permettre aux non-experts de les utiliser facilement. S'investir dans l'utilisation de ces technologies représente donc un risque qu'il est parfois difficile de prendre.

Alternatives plus accessibles : en contrepartie de ce qui est décrit ci-dessus, il est souvent possible de réaliser un projet en utilisant des technologies plus traditionnelles, telles que les systèmes de gestion de bases de données relationnelles, pour lesquels une expérience, de l'aide et un certain niveau de support sont disponibles. Les avantages cités plus haut ne

seront de fait pas réalisés, mais leur importance comparé aux avantages de l'utilisation d'outils établis reste difficile à communiquer aux collaborateurs s'intéressant aux aspects techniques du projet (comme par exemple le service informatique d'un département d'une université).

Coût : beaucoup des outils utilisés pour construire des graphes de connaissances ou pour utiliser les graphes de connaissances dans des applications sont libres et gratuits. Le coût devrait donc être considéré comme un avantage. Il faut néanmoins, comme cité ci-dessus, comparer ce coût à l'utilisation d'outils déjà établis est disponible dans l'équipe du projet. S'ajoute à cela non seulement le coût de déploiement (serveur, maintenance) mais aussi le coût en ressources humaines : il est souvent en effet nécessaire d'acquérir les compétences nécessaires à l'utilisation de ces technologies.

Le but ici est de confronter cette perception à l'expérience concrète de plusieurs projets dans des domaines variés. Les sections suivantes présentent les trois projets, et sont suivies d'une discussion sur les avantages réellement réalisés et les difficultés rencontrées dans ces trois projets, pour conclure sur le besoin d'évoluer les technologies pour lever certaines de ces difficultés et mieux mettre en avant les avantages.

3. Le projet LEDA : l'enfer et les pécheurs en Crète

Le projet LEDA [3] s'intéresse aux représentations de l'enfer dans les églises de Crète. L'intérieur de ces églises est traditionnellement recouvert de fresques murales représentant différents aspects de la religion : des personnages, des scènes bibliques, etc. La localisation de ces fresques au sein de l'église est significative, et le projet s'intéresse tout particulièrement à celles représentant l'enfer, les pécheurs et les tortures qui leur sont infligées. Un des objectifs du projet est de répertorier ces représentations pour comprendre quelles sont les conventions de représentation de l'enfer, de scènes spécifiques, des péchés, en fonction de l'église, de la région, etc. (voir [9], l'ouvrage en deux volumes publié à l'issue du projet).

L'utilisation des graphes de connaissances est ici liée à un besoin de représentation évolutive des informations et à la facilitation de la navigation dans ces informations. En effet, les chercheurs impliqués dans le projet souhaitaient avoir un portail, accessible à l'équipe de recherche et à d'autres, de façon à pouvoir explorer les milliers de photographies collectées et facilement obtenir, par exemple, toutes les représentations de la même scène ou du même péché. Il devait être possible de filtrer les représentation par église, région, localisation au sein de l'église, péché ou scène et la représentation de la localisation devait être précise : il devait clairement apparaître la partie de l'église dans laquelle se trouvait la fresque, sur quel élément architectural, à quelle hauteur, à côté de quelles autres fresques, etc. D'autres informations, telles que les fresques que représente chaque photographie, devaient aussi être présentes.

Une des raisons pour lesquelles ce projet s'est tourné vers les graphes de connaissances est lié à l'accessibilité, la flexibilité et l'évolutivité de ce mode de représentation. En effet, l'annotation et la classification des fresques et de photographies étaient préalablement réalisées au travers d'une base de données Microsoft Access (technologie connue et maîtrisée par certains membres

3. https://ledaproject.org.uk

de l'équipe). Il est apparu clairement néanmoins que la mise à disposition sur le Web d'une telle base de données serait difficile. D'une façon plus importante, un des problèmes de ce mode de représentation est qu'il était nécessaire de fixer le modèle des données à priori, alors que l'équipe n'avait pas encore finalisé les types d'explorations et d'interrogations qu'ils seraient amenés à faire au cours de leur recherche. Finalement, même si les données elles mêmes n'étaient pas de très grande taille, leur richesse (les caractéristiques potentiellement renseignées pour chaque fresque sont nombreuses) et leur hétérogénéité (beaucoup de données incomplètes, avec des valeurs variées, dans plusieurs langues) rendait la création d'une base de données "classique" et sa mise à disposition difficile.

Afin d'utiliser les avantages des technologies du Web sémantique pour pallier ces problèmes, un processus collaboratif et incrémental a été mis en place, où la création d'une ontologie pour la représentation des fresques et des objets liés, la création du portail pour explorer les données et la réflexion sur les interrogations possibles étaient considérées en parallèle et influaient les uns sur les autres. Une des difficultés liées au projet est que les chercheurs en iconographie avaient peu de connaissances en technologie, et ne s'y intéressaient pas *a priori*. Le projet s'appuyait donc sur un dialogue entre ces chercheurs et les informaticiens où différentes itérations de l'ontologie et du portail étaient discutées pour progressivement arriver à un consensus sur les fonctions et le mode de représentation requis. Un aspect intéressant de ce projet est aussi qu'une méthode d'entrée de données utilisant des tableurs, ensuite automatiquement traduits sous la forme de graphes de connaissances, a été mise en place, permettant aux chercheurs de contrôler le contenu du graphe de connaissances sans avoir à s'investir dans l'utilisation des technologies et outils associés. Le résultat de ce processus, le portail d'exploration des informations iconographiques sur l'enfer dans 92 églises de Crète, est représenté figure 1.

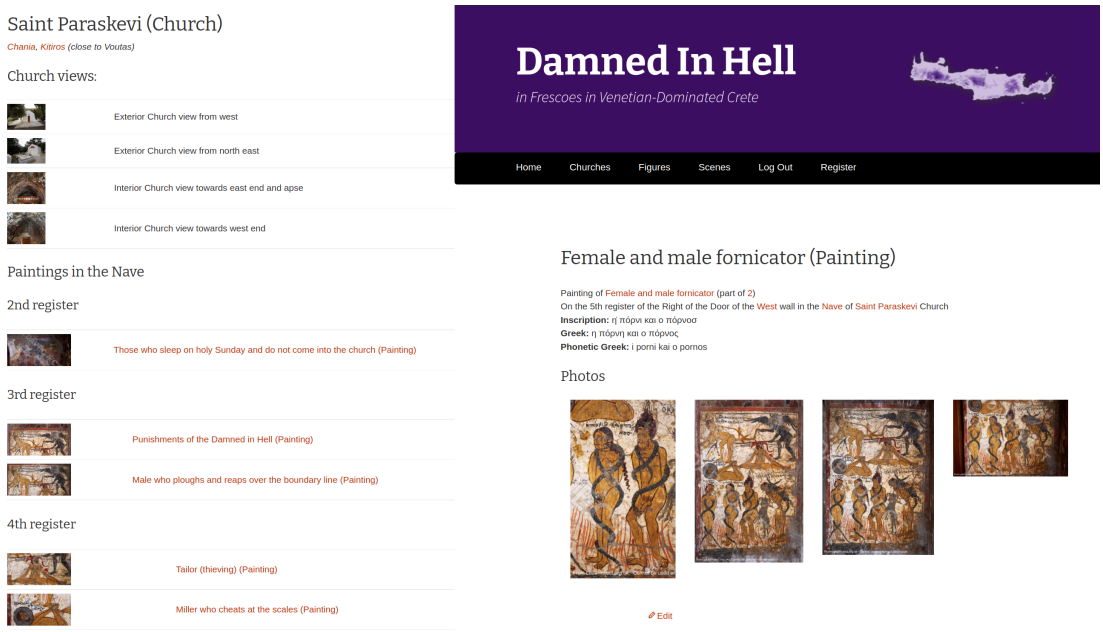


Figure 1 : Le portail du projet LEDA. Vue d'une église (à gauche) et d'une fresque (à droite).

4. Le projet *Listening Experience Database*

The Listening Experience Database (LED) [4] est un projet ambitieux dont le but est de collecter des milliers d'expériences d'écoute de la musique avec autant d'information sur les personnes impliquées, la musique, son contexte et les sources des descriptions que possible (voir [10, 11, 12]). Comme le projet précédent, une des raisons de l'utilisation des graphes de connaissances dans ce projet est la flexibilité et l'évolutivité de la représentation. En effet, les informations à collecter sont très riches, incluant des éléments spatio-temporels (le lieu et le temps de l'écoute de la musique, mais aussi de son exécution), bibliographiques, musicales, socio-économiques, etc.

De ce fait, de façon similaire au projet ci-dessus, les processus de développement de l'ontologie du projet et des outils de navigation dans les graphes de connaissances créés se sont déroulés incrémentalement, au travers d'un dialogue entre les développeurs et les équipes de recherche. De plus, on retrouve ici l'avantage associé aux données liées puisque les graphes créés se connectent à des sources d'information de référence telles que la *British Library* ou VIAF [5].

Deux autres points importants, que l'on peut voir comme des éléments de difficulté, étaient aussi à prendre en compte dans la réalisation de ce projet sur la base de graphes de connaissances : 1- certain éléments d'information était vagues, et 2- les descriptions d'expériences d'écoute de la musique devait être renseignées par les utilisateurs (variés) de la plate-forme (*crowdsourcing*). En effet, les descriptions d'expériences d'écoute (généralement un paragraphe de texte) proviennent de sources variées, incluant des correspondances, journaux personnels, etc. Par nature, ces sources ne contiennent pas toujours toutes les informations requises, mais peuvent aussi contenir des informations imprécises concernant par exemple le moment de l'écoute (*un mardi après-midi, en automne, dans les années 1920*) ou son lieu (*dans le train entre Paris et Lyon*). Pour permettre ce genre de représentations, en évitant au maximum de perdre de l'information, il a donc été nécessaire de construire des structures ontologiques riches autour de notions simples telles que le temps ou la localisation.

L'autre élément, que les données incluses devait provenir d'utilisateurs variés, a posé un certain nombre de difficultés. Encore plus que dans le projet précédent, tout d'abord, cela supposait de permettre l'édition de données d'une façon qui masque la technologie sous-jacente. Cela a été réalisé au travers d'un ensemble de formulaires dont beaucoup de champs sont optionnels et où l'entrée d'information est aussi guidée que possible. Ces formulaires permettent en particulier de réutiliser des éléments entrés par d'autres, réduisant ainsi l'effort requis et évitant les incohérences entre les entrées de différents utilisateurs.

Cet aspect de gestion des incohérences (et des erreurs) représente aussi un élément majeur de ce projet. Plusieurs utilisateurs peuvent renseigner des informations sur les mêmes personnes, musiques, livres, etc. Il se peut que des erreurs se glissent dans les contributions ou même que différents contributeurs aient différentes opinions. Pour permettre une gestion efficace de ces difficultés, un mécanisme de validation a été mis en place où chaque contributeur possède son propre graphe de connaissances, et seulement les éléments de ce graphe de connaissances qui ont été approuvés par un membre du projet sont inclus dans le graphe de connaissances général et publique du projet.

4. https://www.listeningexperience.org/
5. *The virtual authority file* - http://viaf.org/

Le résultat de ces développements est un portail (voir figure 2) qui inclut à l'heure actuelle près de 12 000 descriptions d'expériences d'écoute et qui permet de rechercher, naviguer et explorer ces descriptions en fonctions de nombreux critères (en plus d'ajouter ses propres descriptions). Ce portail permet aux chercheurs de se focaliser par exemple sur certaines périodes, certains lieux, certains genres de musique, ou certains contextes d'écoute, et d'obtenir des informations riches sur les expériences répondant à ces critères. En fournissant ainsi une plate-forme pour l'enregistrement et l'exploration de ces expériences d'écoute, LED offre ainsi un support de recherche pour de nombreux chercheurs, comme en témoignent les actes édités en ligne des deux conférences déjà organisées sur le sujet [6]. La création de cette base de descriptions d'expériences d'écoute de la musique a en plus permis le développement de nouvelles applications intelligentes, telles que *FindLEr* [7] qui permet de retrouver dans des textes quelconques des passages qui ressemblent à des descriptions d'expériences d'écoute de la musique.

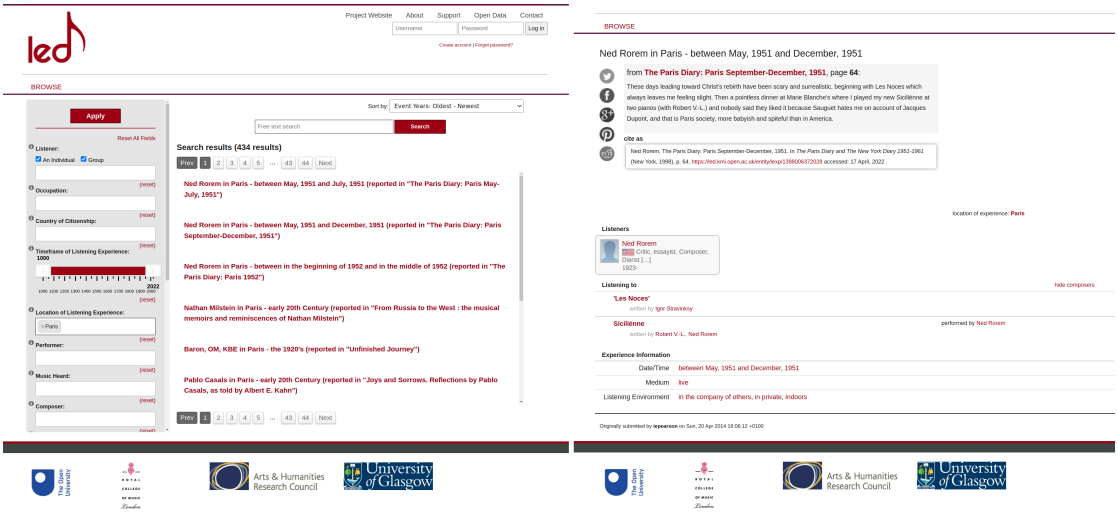


Figure 2 : Le portail du projet LED. Interface de recherche d'expériences d'écoute de la musique (à gauche - recherche d'expériences localisées à Paris) et description d'une expérience d'écoute (à droite).

5. Correspondances entre John Norris, Mary Astell et Damaris Masham

L'étude de textes philosophiques est une activité par nature complexe et sujette à interprétation. C'est d'autant plus le cas quand ces textes forment des correspondances entre plusieurs auteurs. Le but de l'outil ArguNest [8], dévelopé dans le cadre de la thèse de Ioanna Kyvernitou (NUI Galway, Irlande) est de permettre de s'abstraire des textes eux-même et de représenter, au travers de graphes de connaissances, la *lecture* de ces textes par le réseau d'arguments et de propositions qu'ils contiennent.

6. http://ledbooks.org/
7. https://led.kmi.open.ac.uk/discovery/findler
8. https://github.com/mdaquin/ArguNest

Le cœur de ce projet est donc une ontologie permettant de représenter ce réseau d'arguments et de propositions. L'analyse d'arguments est une discipline établie et de nombreux modèles de représentation de réseaux d'arguments existent. On s'intéresse ici en particulier aux lettres échangées par John Norris, un théologue reconnu de son époque, et Mary Astell et Damaris Masham, qui remettent en cause certains des raisonnements exposés dans ses ouvrages. Un aspect important dans la représentation ontologique de ces échanges était donc la capacité de référencer des arguments déjà établis, et donc de séparer la notion abstraite d'argument de sa matérialisation dans les textes. En effet, le même argument peut être réutilisé, re-discuté ou mentionné plusieurs fois dans les textes, sous plusieurs formes différentes. L'objectif était donc de pouvoir représenter une lecture particulière des textes à deux niveaux d'abstraction : les annotations des textes comme représentant des arguments et des propositions, et comment ces arguments et ces propositions, en tant qu'entités abstraites, sont liés entre eux.

L'outil créé sur la base de cette ontologie (ArguNest, voir figure 3) est donc essentiellement un outil d'annotation de textes permettant d'éditer un graphe de connaissances avec des informations sur ces deux niveaux d'abstraction. Les textes sont représentés de façon à permettre d'identifier des expressions d'arguments et de propositions et de décrire ces arguments et ces propositions. Une fois ces arguments et propositions identifiés, une autre partie de l'interface permet de créer des relations entre ceux-ci. Le résultat de l'utilisation de cet outil est donc un graphe de connaissances qui correspond à la lecture faite par l'utilisateur de ces textes. L'outil est développé de façon générique et peut donc être utilisé sur n'importe quel texte philosophique.

Un aspect particulièrement intéressant ici est de voir comment l'utilisation des technologies liées aux graphes de connaissances affecte la méthodologie de recherche utilisée et la façon de travailler sur l'étude de ces textes. En effet, la création du graphe de connaissances peut être vue comme une activité de recherche, et le graphe devient lui-même un objet d'étude, explorable et analysable, en plus des textes. De plus, cela facilite l'échange en rendant comparable les graphes de connaissances obtenus des lectures de différents utilisateurs. L'outil a d'ailleurs été testé dans un contexte d'enseignement, ou des étudiants en littérature et philosophie s'appuient sur l'annotation de textes comme moyen d'étude, et peuvent ensuite échanger sur leur compréhension de textes complexes sur la base de la comparaison des graphes obtenus.

6. Discussion : les avantages réels et les difficultés communes

Il n'est bien sûr pas dans l'intention de cet article de prétendre que les trois exemples de projets présentés ci-dessus sont représentatifs de projets typiques utilisant les graphes de connaissances pour les humanités numériques. Néanmoins, même s'ils ont de nombreux points communs, ceux-ci sont suffisamment différents pour que l'on puisse en tirer quelques leçons sur les avantages à utiliser des graphes de connaissances et certains problèmes qu'il reste à pallier.

En effet, un des éléments communs à ces trois projets est que les avantages obtenus ne sont pas nécessairement alignés avec ceux attendus. En effet, par exemple, seulement LED a été amené à développer une application intelligente et a réellement utilisé les liens avec d'autres sources d'information disponibles sous la forme de graphes de connaissances. Même pour LED, ces deux aspects ne sont pas réellement centraux et restent anecdotiques. De la même façon, l'accès à l'information reste important pour ces projets, mais sous une forme différente du

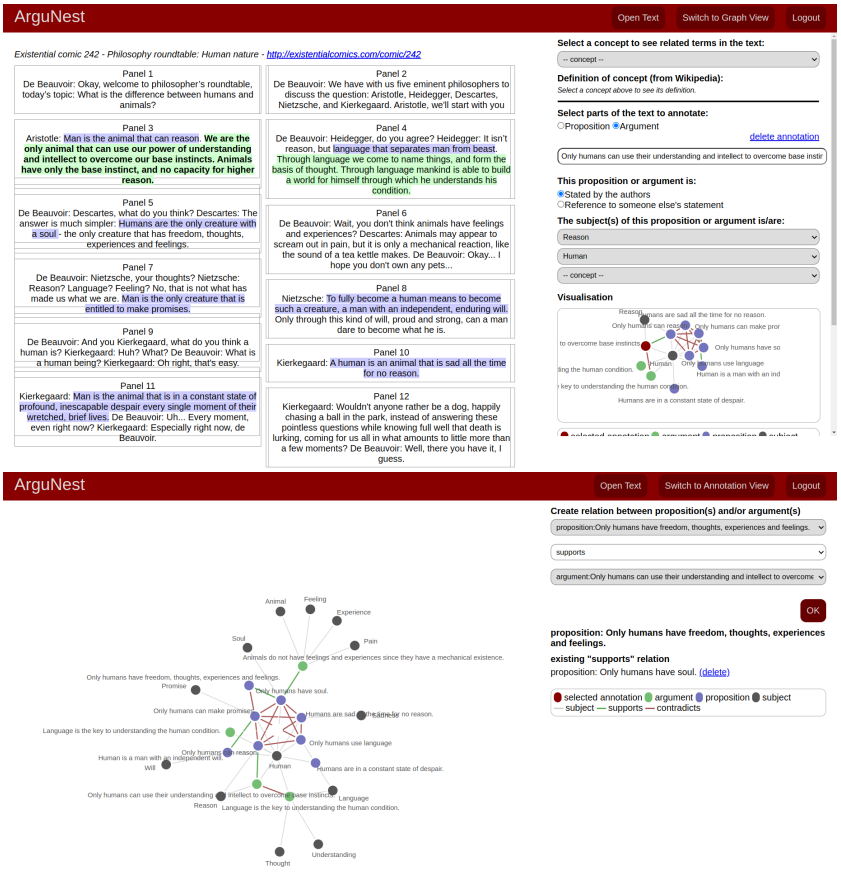


Figure 3 : L'outil ArguNest. Interface d'annotation de textes (haut) et de mise en relation des arguments et propositions (bas).

graphe de connaissances pur. Dans chacun de ces cas une interface masquant la technologie impliquée a été développée. Ces interfaces bénéficient de la mise sous forme de graphes de connaissances des informations à traiter, mais sont réalisées en plus pour faciliter l'exploration par des utilisateurs non familiarisés avec les principes du Web sémantique. Cela a, de fait, aussi une implication sur les attentes en termes de visibilité. D'autres projets autour du patrimoine culturel par exemple mettent cet aspect plus en avant, alors que dans le cas des trois projets considérés ici, l'apport de la création de graphes de connaissances à la visibilité du projet et de la recherche reste moindre.

Bien sûr, il ne faut pas conclure de ce qui est écrit ci-dessus que l'utilisation des technologies du Web sémantique n'a pas d'avantages. Tout projet en humanités numériques est, par nature, pluridisciplinaire, mais un des aspects essentiels de ces trois projets est qu'ils étaient tous les trois fondés sur une collaboration forte entre chercheurs en sciences humaines et chercheurs/développeurs en informatique. Comme montré plus haut, si on se focalise purement sur les éléments liés à la technologie, le fait que les graphes de connaissances permettent une

représentation flexible et évolutive semble être le point le plus important. En réalité, on peut aussi voir ce point technique comme étant fondamental au point positif le plus important dans ces trois projets : que la création sans contrainte technique forte d'une conceptualisation du domaine considéré représente une tâche pivot entre les disciplines, permettant d'en aligner les notions, les vocabulaires et les attentes.

En effet, dans chacun des trois projets, l'élément à la base de la collaboration et ayant permis de la construire était la création d'une ontologie qui réponde aux besoins du projet, et qui devait être le pilier central des outils et systèmes développés. La construction de cette ontologie doit, par nature, être une activité collaborative avec en son centre l'établissement d'une conceptualisation consensuelle des notions du domaine. Cette représentation doit être encodable dans les formalismes de représentation utilisés. Une des raisons de se tourner vers les technologies du Web sémantique et le graphe de connaissances est, comme déjà exprimé plus haut, qu'ils permettent des représentations riches, en incluant notamment des éléments incomplets et incertains. Les projets présentés ici ont débuté avec une vision vague et peu structurée de ce qui devait devenir cette conceptualisation. Là où d'autres types de technologies auraient nécessité de forcer une structure fortement contrainte sur la représentation des concepts du domaine et des informations associées dès le début du projet, la construction itérative d'une ontologie ici a permis une clarification progressive de ces éléments, l'explicitation des problèmes et des aspects spécifiques de chaque projet et la mise en place d'une vue partagée des éléments du projet au travers d'un dialogue plus équilibré entre le domaine de recherche concerné et la technique. Le résultat de ce processus est un artefact informatique qui non seulement va permettre de structurer le reste des développements technologiques dans le projet, mais qui va aussi encapsuler une vision commune du cœur du sujet entre les participants.

Que le plus significatif des avantages dans l'utilisation des graphes de connaissances soit au niveau de la conceptualisation et du dialogue entre les disciplines est aussi lié à un inconvénient majeur de ces technologies. En effet, malgré ce qui est écrit plus haut, les outils pour construire des ontologies, pour éditer des graphes de connaissances et pour naviguer au sein de ces graphes de connaissances restent complexes et obscures pour les non-spécialistes. Les conceptualisations initiales se font, par conséquent, souvent sur papier, sur des tableaux blancs ou sur la base d'outils non-dédiés, mais mieux maîtrisés ou maîtrisables par les experts du domaine de recherche. L'utilisation de tableurs en ligne dans le projet LEDA est un parfait exemple de ce type de difficultés qui nécessite de mettre en place des représentations intermédiaires que la partie plus technique du projet devra ensuite transformer en une représentation compatible avec l'utilisation des graphes de connaissances. Un autre exemple est le développement de l'outil OWBO [9] directement motivé par l'expérience dans ces trois projets. En effet, les outils de construction d'ontologies tels que Protégé [10] sont très peu adaptés à la phase initiale de structuration d'une ontologie et ne permettent pas facilement aux experts du domaine, non-spécialistes des technologies du Web sémantique, d'être directement impliqués dans la construction de l'ontologie. Une représentation séparée est souvent construite, par exemple sur un tableau blanc, laissant aux informaticiens le soin de les retranscrire dans Protégé. L'idée d'OWBO est

9. https://github.com/mdaquin/OWBO
10. https://protege.stanford.edu/

de fournir une version épurée, simplifiée et partageable de la création d'une ontologie initiale, qui s'apparente à l'utilisation d'un tableau blanc, et qui peut être transférée directement dans des outils tels que Protégé pour être affinée.

Finalement, un des désavantages les plus importants de l'utilisation des graphes de connaissances, peu visible dans la description des projets dans cet article mais tout de même très présent, est lié au manque de maturité des outils et des systèmes utilisés, et à leur pérennité. Beaucoup de ces systèmes sont développés dans des équipes de recherche ayant peu de moyens pour garantir leur fonctionnement et pour les mettre à jour autant qu'il peut être nécessaire. Comme discuté au début de cet article, alors que leur coût est moindre (ils sont souvent gratuits), le problème que cela amène est qu'il devient difficile de maintenir les graphes de connaissances et les applications construits sur la base de ces outils et systèmes. Le coût ici se retrouve concentré dans le temps des spécialistes en technologies du Web sémantique requis pour non seulement collaborer à la création de ces applications, mais aussi pour s'assurer que celles-ci continuent de fonctionner à long terme.

7. Conclusion

Dans cet article est décrit succinctement trois expériences de projets en humanités numériques utilisant des graphes de connaissances dans trois domaines différents : l'iconographie, l'histoire de la musique et la littérature/philosophie. Alors que pour les chercheurs impliqués dans le développement de ces technologies, les avantages de l'utilisation des graphes de connaissances peuvent paraître évidents, la réalité de leur mise en place et du développement d'applications les utilisant dans ce type de domaines n'est pas toujours alignée en pratique avec ce qui est attendu. Au travers de ces trois projets, il est possible de mieux comprendre comment, au-delà des aspects purement techniques, un des points essentiels des graphes de connaissances est qu'ils fournissent un modèle du domaine, complètement partagé et offrant une vision commune de ce qui est le cœur du projet. Néanmoins, pour que ces projets réussissent, il est nécessaire que cette conceptualisation soit réalisée au sein d'une vraie collaboration pluridisciplinaire entre les experts du domaine, qui possèdent la connaissance requise et les attentes liées au projet, et les spécialistes des technologies associées, capable de mettre en place les modèles de connaissances construits. Cela nécessite une forte implication de la part de ces spécialistes, au moment du développement et par la suite, d'autant que les outils disponibles actuellement ne sont pas vraiment conçus pour faciliter la collaboration ou la maintenance à long terme des graphes de connaissances et des applications les utilisant.

Remerciements

L'auteur tient à remercier les membres des projets LEDA, LED et ArguNest pour leur collaboration et contributions sur la base desquelles cet article a été écrit.

Références

[1] S. Ji, S. Pan, E. Cambria, P. Marttinen, S. Y. Philip, A survey on knowledge graphs : Representation, acquisition, and applications, IEEE Transactions on Neural Networks and Learning Systems (2021).

[2] G. Antoniou, F. Van Harmelen, A semantic web primer, MIT press, 2004.

[3] E. Hyvönen, Using the semantic web in digital humanities : Shift from data publishing to data-analysis and serendipitous knowledge discovery, Semantic Web 11 (2020) 187–193.

[4] S. Middle, Investigating Linked Data usability for Ancient World research, Ph.D. thesis, The Open University, Miton Keynes, UK, 2022.

[5] S. Staab, R. Studer, Handbook on ontologies, Springer Science & Business Media, 2010.

[6] D. Mouromtsev, M. d'Aquin, Open data for education : Linked, shared, and reusable data for teaching and learning, volume 9500, Springer, 2016.

[7] M. d'Aquin, E. Motta, The epistemology of intelligent semantic web systems, Synthesis Lectures on the Semantic Web : Theory and Technology 6 (2016) 1–88.

[8] R. Wenz, Linked open data for new library services : the example of data. bnf. fr., Linked open data for new library services : the example of data. bnf. fr. (2013) 403–416.

[9] A. Lymberopoulou, Hell in the Byzantine World : A History of Art and Religion in Venetian Crete and the Eastern Mediterranean, Cambridge University Press, 2020.

[10] A. Adamou, M. d'Aquin, H. Barlow, S. Brown, Led : curated and crowdsourced linked data on music listening experiences (2014).

[11] S. Brown, H. Barlow, A. Adamou, M. d'Aquin, The listening experience database project : Collating the responses of the" ordinary listener" to prompt new insights into musical experience., International Journal of the Humanities : Annual Review 13 (2015).

[12] A. Adamou, S. Brown, H. Barlow, C. Allocca, M. d'Aquin, Crowdsourcing linked data on listening experiences through reuse and enhancement of library data, International Journal on Digital Libraries 20 (2019) 61–79.

Expérimentations sémantiques autour de la Chanson de Roland

Semantic experiments around the Chanson de Roland

Jacques Ducloy[1,*], Thierry Daunois[2] and Isabelle Turcan[2]

[1]*Université Paris 8, Laboratoire Paragraphe, F-93200 Saint-Denis, France*
[2]*Université de Lorraine, F-54000 Nancy, France*

Abstract

This article introduces a hypertext digital library on the *Chanson de Roland*. It collects manuscripts, critical editions, translations, research articles and musical scores. It is both a workspace for specialists in humanities and a source of information for a curious but non-specialist readers. Articles and manuscripts are republished in hypertext mode with a common semantic structure. The current demonstrator is using 3 manuscripts (Oxford, Paris, Châteauroux) and critical editions (Francisque Michel, Léon Gautier, Edmund Stengel, Joseph Bédier). Two applications are presented. Specialists can work on part of the Paul Meyer collection. Curious amateurs, for example choristers, can explore the context of a secular oratorio by Gilles Mathieu. This diversity implies taking into account various digital approaches which are experimented here with Semantic MediaWiki, and XML engineering. The generalization of this approach is studied.

Keywords

Chanson de Roland, époque carolingienne, wiki sémantique, manuscrits, bibliothèque numérique, édition critique, semantic mediaWiki, musique

Avant-Propos

Une version numérique augmentée (page de discussion, liens hypertextes et sémantiques) est disponible sur le site Wicri/Chanson de Roland [1].

1. Introduction

Le 15 aout 778, de retour d'Espagne, Charlemagne perd son arrière-garde, tombée, à titre de représailles, sous le feu des troupes des seigneurs basques dont il a attaqué les possessions. Lors de la bataille de Roncevaux, l'arrière-garde est écrasée, provoquant la mort de nombreux braves de l'entourage de Charlemagne, dont celle de Roland, préfet de la Marche de Bretagne. Ce fait d'armes a inspiré des cantilènes, des récits et une chanson de geste, la *Chanson de Roland*. Ce poème épique a été déclamé dans toute l'Europe par des jongleurs et des troubadours. Quelques

. *Workshop on Digital Humanities and Semantic Web*
∗. Corresponding author.
. Jacques.Ducloy@univ-lorraine.fr (J. Ducloy); thierry.daunois@univ-lorraine.fr (T. Daunois); isabelle.turcan@univ-lorraine.fr (I. Turcan)

1. https://wicri-demo.istex.fr/Wicri/Europe/ChansonRoland/fr/index.php/Article_Humanum_Nancy_2022.

manuscrits ont survécu et font l'objet d'une abondante production littéraire depuis le XIX^{e} siècle.

Mais ces écrits n'étaient pas toujours accessibles facilement. Les manuscrits étaient enfermés dans des bibliothèques dispersées (Oxford, Paris, Venise, Châteauroux...). Les ouvrages étaient souvent édités avec une diffusion modeste à destination d'un public d'érudits comme les élèves de l'École nationale des chartes, à côté d'éditions grand public. Le numérique permet aujourd'hui d'accéder à cette littérature. Mais cette dernière est toujours dispersée sur de multiples sites qui ont chacun leurs modes d'accès.

Le fonds Paul Meyer de l'Université de Lorraine contient un document particulièrement intéressant : une édition de 1869 de « *La Chanson de Roland, ou de Roncevaux, du XII^{e} siècle* » de Francisque Michel [1], annotée par Paul Meyer. Celui-ci a ainsi effectué un travail préparatoire à une de ses publications [2]. Pour confronter les points de vue des deux auteurs aux manuscrits originaux, des centaines de laisses [2], avec leurs transcriptions et leurs traductions, sont manipulées. Ce problème est apparu comme particulièrement pertinent pour le réseau Wicri, un projet sur les bibliothèques qui gèrent des collections de documents hypertextes.

Par un concours de circonstances, nous avions travaillé avec un musicien, Gilles Mathieu, qui a composé une suite musicale à partir des mêmes manuscrits, mais sur la base d'une autre traduction [3]. Cette composition amène un nouveau point de vue qui enrichit cet ensemble. Elle ouvre également le site à un nouveau public, les choristes amateurs, qui sont des lecteurs curieux mais pas forcément érudits. Cette contrainte implique notamment de rééditer d'autres documents plus explicatifs.

Nous avons donc décidé de constituer une bibliothèque numérique spécialisée autour de la *Chanson de Roland*. Ce projet a déjà été présenté, dans sa phase de démarrage, avec un éclairage de valorisation du patrimoine écrit [4]. Nous présentons ici les premières avancées et un éclairage sur les aspects sémantiques.

Après une description des relations sémantiques dans le réseau Wicri, nous détaillerons l'organisation retenue pour les manuscrits et leurs traductions. Puis nous montrerons les premières réalisations autour du fonds Paul Meyer et de la suite musicale.

2. Les relations sémantiques dans le réseau Wicri

Le projet Wicri (Wikis pour les communautés de la recherche et de l'innovation) [3] a été créé en 2008. Pour les communautés de la recherche, il explore de nouvelles approches numériques en s'inspirant des mécanismes et pratiques mises en œuvre dans Wikipédia dont le moteur (MediaWiki) favorise un développement collectif et incrémental.

Un premier réseau d'une dizaine de wikis avait été expérimenté pour valoriser les résultats de la recherche en Lorraine autour des sciences et du génie de l'environnement. Une coopération avec le Loria a ouvert l'usage des extensions sémantiques (Semantic MediaWiki). Elle avait permis de modéliser les équipements financés par le Contrat de Projets État Région (CPER). Plus tard, un système d'information évolutif sur les projets européens en Lorraine a été développé.

2. Une laisse est un couplet composé de vers ayant la même assonance (voir plus loin).
3. https://wicri-demo.istex.fr/Wicri/Wicri/fr/index.php?title=Accueil.

Pour ces actions, un modèle initialisé sur l'ancien site Semantic Web [4] a été adapté pour décrire les systèmes de recherche, notamment autour des colloques.

Ce modèle a été utilisé sur la plupart wikis, et notamment, pour ceux dédiés aux communautés de colloques (notamment CIDE [5] ou H2PTM [6]). La figure 1 montre, en 2021, l'ensemble des wikis communs en français du démonstrateur Wicri. Ils sont généralement associés à un wiki en anglais (et parfois en allemand) [7].

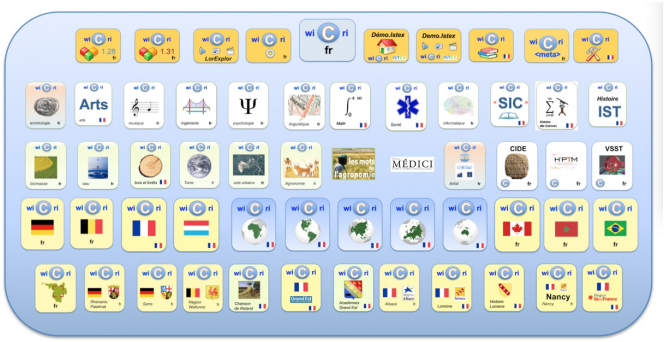


Figure 1: Le réseau Wicri en 2021.

2.1. Un réseau de bibliothèques sur base encyclopédique

Après cette première étape sur la valorisation des résultats de la recherche, deux séries d'études ont été menées.

Pour les sciences relevant de l'ingénierie, de l'environnement et de la santé des résultats très intéressants ont été obtenus avec l'analyse statistique de corpus bibliographiques.

Un financement ISTEX [8] a permis de créer plus d'une centaine de serveurs d'explorations. Un tel outil traite des milliers de références hétérogènes (ISTEX, Pascal, HAL, PubMed). Il est créé à l'aide d'une boîte à outil XML nommée Dilib [6] dont la conception initiale a été réalisée à l'INIST [7].

Dans sa version initiale, un serveur d'exploration était généré par des commandes Unix avec un paramétrage complexe et sans accès au texte intégral. Le wiki est maintenant utilisé pour le paramétrage, la visualisation de résultats significatifs, et la curation des données. Plus précisément, les relations sémantiques utilisées dans la valorisation des innovations demandent une grande précision dans l'identification des données. Celles-ci seront utilisées pour définir les

4. Ce modèle est maintenant soutenu sur le site OpenResearch.org [5].
5. Colloque international sur le document électronique.
6. Hypertexte et hypermédia Produits, Outils et Méthodes.
7. Malheureusement, avec des porteurs alors « proches de la retraite » les travaux ont été poursuivis, mais avec des moyens humains limités à un retraité (et le financement d'un demi-poste d'ingénieur d'études pendant la durée du programme ISTEX). Paradoxalement, cet état de fait est significatif pour apprécier les performances de cette approche.
8. ISTEX (Initiative d'excellence de l'Information Scientifique et Technique) projet retenu dans le cadre du programme « Investissements d'Avenir ».

règles de curation. Par exemple, l'Université de Groningue est localisée à Groningue dans une région éponyme des Pays-Bas. Le modèle sémantique contient alors des triplets tels que :

Groningue (ville) `A pour région:` Groningue (province)

Les règles de curation vont utiliser cette nomenclature pour inférer des mentions géographiques à partir de la mention d'une université dans une affiliation. Voici un exemple qui peut être activé avec « *Rijksuniversiteit Groningen* ».

Université de Groningue	Rijksuniversiteit Groningen ; University of Groningen	country : Pays-Bas ; region : Groningue (province) ; settlement @type=city : Groningue (ville)

Figure 2: Un exemple de règles de curation exprimées avec des tables MediaWiki.

Pour les humanités, cette approche donne des résultats plus limités. En effet, des sources de données très structurantes comme Pascal ou PubMed ne sont plus utilisables. De plus, les corpus ISTEX sont souvent constitués de « books review » qui traitent de sujets variés rassemblées dans un même document numérique. Les résultats statistiques donnent alors des corrélations aberrantes [9]. En revanche des résultats très pertinents ont été obtenus avec des rééditions hypertextes (et sémantisées) de documents anciens (libres de droit).

Le premier résultat significatif a été obtenu avec un ouvrage sur le Palais ducal de Nancy [10]. À partir d'un facsimilé en mode « image + OCR » sur Gallica, nous avons notamment montré comment transformer en hypertexte une gravure de fin de volume (Fig. 3). Elle contenait des liens, matérialisés par des lettres, qui pointaient vers un hypertexte de paragraphes descriptifs qui eux-mêmes renvoyaient à des pages du livre.

Cette approche a été généralisée dans les articles scientifiques pour quelques colloques. Puis, dans un wiki sur la musique, des entrées du dictionnaire de Jean-Jacques Rousseau ont été réédités avec la possibilité d'écouter les partitions. Ainsi, un dictionnaire devient alors un document totalement hypertextuel (là où par exemple Gallica conserve une vision linéaire).

En appliquant cette approche au dictionnaire TLF [11], les auteurs cités deviennent alors des points d'entrée potentiels. Ainsi, sur un wiki dédié à la santé nous avons pu associer à une réédition d'un ouvrage de Claude Bernard [12] de nombreux articles du TLF.

Dans le réseau Wicri, un site wiki devient donc une bibliothèque spécialisée qui utilise une base encyclopédique pour mettre en relation des ouvrages réédités. Il devient également un

9. Par exemple, un corpus ISTEX de 1 500 documents sur le compositeur William Byrd donne 360 mentions de l'Islam (dont aucune n'est significative).

10. https://wicri-demo.istex.fr/Wicri/Europe/France/GrandEst/Lorraine/Nancy/fr/index.php/Le_Palais_ducal_de_Nancy_(1852)_Lepage.

11. Trésor de la langue française, dictionnaire du CNRS, Nancy, 1971-1994, publication papier ; 2004, édition numérisée sur cédérom ; consultation en accès libre sur : https://www.atilf.fr/ressources/tlfi.

12. https://wicri-demo.istex.fr/Wicri/Sante/fr/index.php/Introduction_médecine_expérimentale_(1865)_Bernard.

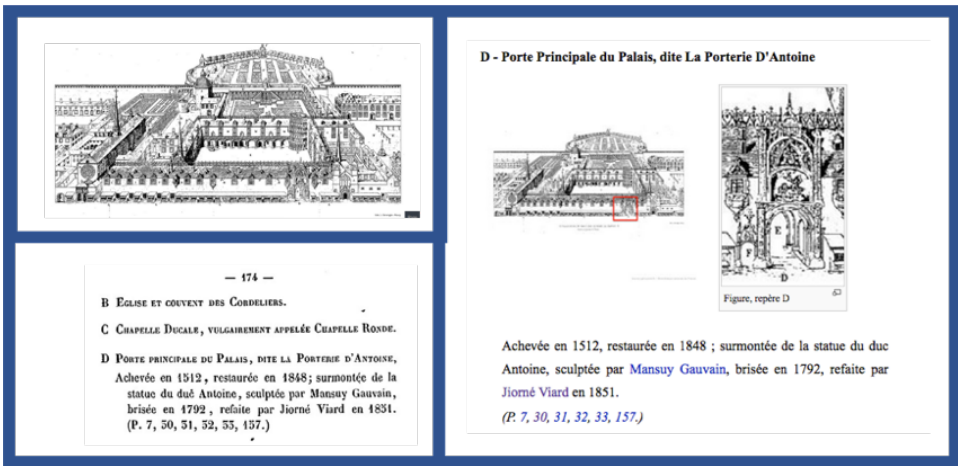


Figure 3: Le Palais ducal : à gauche la gravure et une rubrique (D) avec des renvois ; à droite, le développé de la rubrique D en hypertexte.

espace de travail, où il est, par exemple, possible de piloter collectivement des explorations de corpus.

2.2. Les relations sémantiques en réseau

Dès le lancement du réseau Wicri la cohérence terminologique et sémantique du réseau a fait l'objet d'investigations [8]. A titre d'exemple simple, la figure 4 montre l'alignement des éléments géographiques entre les wikis du réseau.

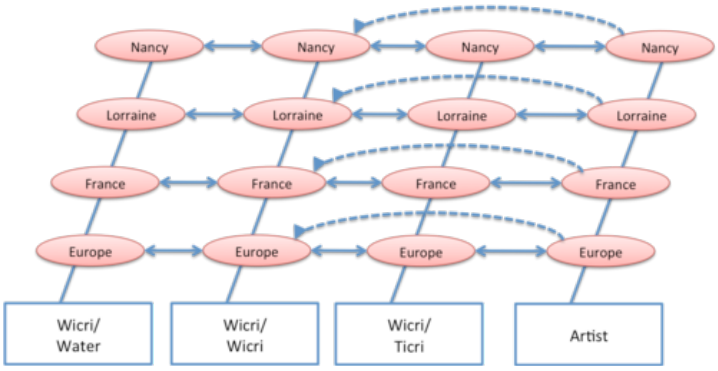


Figure 4: Alignement des relations géographiques entre les wikis.

Cette cohérence est basée sur un alignement sur le Web sémantique. Plus précisément les noms de page sur les wikis sont, si possible, les mêmes, que ceux de Wikipédia. Pour favoriser cet alignement, de nombreux modèles (par exemple la « palette des régions administratives de

France ») sont importés de Wikipédia et éventuellement adaptés. Ces modèles communs sont gérés sur un des wikis du réseau (Wicri/Base). Ils sont regroupés en collections pour favoriser des opérations d'exportation (depuis Wicri/Base) vers les wikis cibles. Actuellement, tous les wikis sont sur le même site physique, et ces actions sont réalisées par des traitements par lots.

Un autre mécanisme, nommé wiki de référence, est également utilisé. Par exemple, l'Université McGill a naturellement Wicri/Canada pour wiki de référence. Lorsqu'une activité significative de cette université est détectée sur un autre wiki, par exemple sur Wicri/Musique, une page spécialisée y est alors créée. Sur celle-ci, un lien interwiki pointe vers la page de référence (sur Wicri/Canada). Enfin, sur ce dernier, un lien est établi vers Wicri/Musique. Ces opérations sont en fait très rapides pour des entités déjà signalées. Cela dit, la création d'un nouveau wiki demande une adaptation du réseau. Par exemple, avant la création de Wiki/Canada, les entités canadiennes étaient sur Wicri/Amérique. Il a donc fallu passer quelques heures pour mettre à jour le réseau de liens [13]. Le maintien de la cohérence du signalement des universités françaises en mutation permanente s'avère nettement plus complexe et montre la nécessité d'une administration terminologique, et surtout éditoriale.

3. Les manuscrits et leurs éditions critiques

Nous venons de présenter la structure d'accueil de l'expérimentation sur la *Chanson de Roland*. Nous allons maintenant introduire les ressources bibliographiques fondamentales de ce sujet : les manuscrits originaux et les éditions critiques associées. Dans une bibliothèque universitaire classique, ce sujet occupe quelques décimètres de rayonnage sous la forme de quelques livres de références (Francisque Michel, Léon Gautier, Joseph Bédier, Joseph Duggan, etc).

Ici, pour permettre des études comparatives, ces quelques livres vont alimenter, à moyen terme, un réseau hypertexte de plusieurs dizaines de milliers d'articles.

3.1. Un corpus riche et varié

3.1.1. Les manuscrits

De la *Chanson de Roland* et de ses transcriptions médiévales, on connait aujourd'hui sept versions, et trois fragments. La version considérée comme la plus ancienne et la plus proche d'un hypothétique « texte initial » est le manuscrit conservé à la Bibliothèque Bodléienne d'Oxford (Digby, 23, f. 1r-72r). Communément daté du deuxième quart du XII^{e} siècle, ce manuscrit a suscité plusieurs dizaines d'éditions modernes, depuis le début du XIX^{e} siècle, a été traduit dans de nombreuses langues, et fait l'objet de plusieurs centaines d'études [14].

Une analyse même sommaire des versions manuscrites de la chanson de geste permet immédiatement de comprendre la situation. Là où le manuscrit d'Oxford compte 4002 vers répartis en 291 laisses (ou couplets), la version Venise 4 — datée du XIII^{e} siècle — en compte 6011, pour 419 laisses, la version de Châteauroux, 8201 vers et 449 laisses, le manuscrit Venise 7 rassemble 8395 vers organisés en 445 laisses. Les manuscrits de Paris, Cambridge et Lyon, pour leur part,

13. Une telle opération pourrait assez facilement être partiellement automatisée par un robot.

14. La consultation de la bibliographie proposée sur le site arlima.net est éclairante sur la richesse et des écrits sur la *Chanson de Roland* : https://www.arlima.net/qt/roland_chanson_de.html.

comptent respectivement 6828, 5695 et 2932 vers, distribués en 375, 354 et 216 laisses. Chaque manuscrit possède sa propre variante linguistique — par exemple, Venise 4 est en italien francisé. Les mécanismes de versification sont variables, de l'assonance à la véritable rime.

Ces manuscrits sont organisés en laisses. Une laisse est une suite de vers avec une unité de versification (assonance sur le manuscrit d'Oxford), et généralement matérialisée par une lettrine (Fig. 5). Dans le manuscrit d'Oxford, elles se terminent par une mention mystérieuse [Aoi], sur laquelle aucune explication ne semble unanimement acceptée [9].

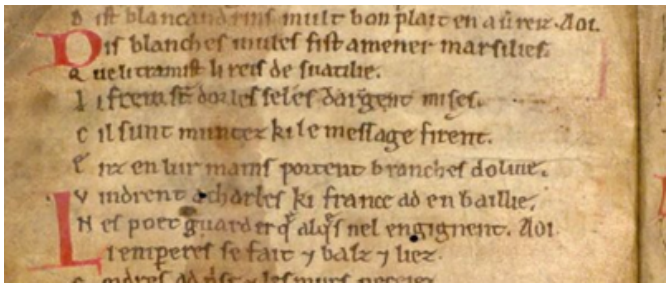


Figure 5: Un enchaînement de 3 laisses, 2 lettrines (D et L) et 2 mentions Aoi en fin de ligne.

Au-delà de la forme poétique chaque laisse contient une partie du récit. Une grande majorité de laisses traitent des même faits (avec cependant des variantes locales) sur les différents manuscrits. Voici par exemple le début de la première laisse dans le manuscrit d'Oxford :

> Carles li Reis, nostre emperere magnes,
> Set anz tuz pleins ad estet en Espaigne
> Cunquist la tere tresqu'en la mer altaigne.

Sur le manuscrit de Châteauroux, ce passage devient :

> Challes li rois à la barbe grifaigne
> Sis anz toz plens a esté en Espaigne,
> Conquist la terre jusque la mer alteigne

Pour les lecteurs non familiers avec la langue romane voici la traduction donnée par Léon Gautier pour le manuscrit d'Oxford :

> Charles le roi, notre grand empereur,
> Sept ans entiers est resté en Espagne :
> Jusqu'à la haute mer, il a conquis la terre.

Les chiffres donnés plus haut sur le nombre de vers et de laisses montrent une très grande variété de situations (ajout ou retrait de vers, éclatement de laisses, etc.).

Ces laisses sont distribuées sur des feuillets avec un découpage basé généralement sur un nombre de lignes par page ou par colonne. Une laisse peut ainsi être à cheval sur plusieurs feuillets.

D'un point de vue informatique, la colonne vertébrale de ce rayonnage numérique est donc une juxtaposition de 2 arborescences (avec ou sans le niveau feuillet) et des relations pas toujours

binaires entre les laisses. Sur cette base, l'interprétation donnée par les philologues introduit un nouveau niveau de complexité.

3.1.2. Divergences entre les transcriptions et éditions critiques

Lorsque l'on commence à vouloir aligner les textes des manuscrits et leurs transcriptions, on constate rapidement des divergences dans la numérotation des laisses. Ainsi, la dernière laisse du texte est numérotée CCXCI chez Joseph Bédier, CCXCIII chez Edmund Stengel, CCXCVI chez Francisque Michel et CCXCVII chez Léon Gautier, alors qu'ils sont censés avoir travaillé sur le même manuscrit de départ (en l'occurrence, le manuscrit d'Oxford).

En effet, certains philologues se réfèrent à la différenciation des laisses à l'aide des lettrines et des marques [Aoi] telle qu'elle est dans le manuscrit d'Oxford. D'autres sont plus attentifs à la versification. Certains ont eux-mêmes commis une erreur de numérotation. D'autres enfin considèrent que le copiste a fait des erreurs qu'il faut rectifier. Le feuillet 43 verso est exemplaire de ce point de vue car il ne contient ni lettrine, ni mention [Aoi]. En revanche, il contient un vers qui marque une charnière essentielle entre deux parties de l'épopée : la mort de Roland.

> Morz est Rollant, Deus en ad l'anme es cels.
> Roland est mort ; Dieu a son âme dans les cieux.

Le manuscrit contient curieusement un point (en guise de lettrine ?), avant ce vers (Fig. 6).

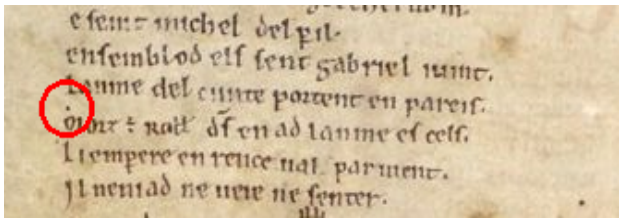


Figure 6: Le verset 43 verso.

Bédier et Gautier considèrent ce vers comme le début d'une nouvelle laisse. Michel en fait la fin de la précédente et Stengel propose une version sans changement de laisse (et donc avec un décalage dans la numérotation).

3.2. Gestion numérique des manuscrits et des éditions critiques

À partir d'investigations menées dans le cadre d'un stage, nous avons confronté le manuscrit d'Oxford avec les versions de Francisque Michel, Léon Gautier et Joseph Bédier.

Dans un premier temps, nous avons demandé à l'étudiant de réaliser un alignement entre le manuscrit d'Oxford et la version de Francisque Michel. Plus précisément, les laisses étaient identifiées (au sens numérique) en utilisant la numérotation de Michel. Malheureusement cette approche était insuffisante pour prendre en compte de façon précise les analyses de Gautier et de Bédier. Nous avons donc décidé de gérer les manuscrits en nous appuyant sur les laisses visibles par un public non forcément érudit, et avec notre propre numérotation.

En même temps, l'exploration des sources a mis en évidence un ouvrage d'Edmund Stengel [10] dans lequel la pagination suit le découpage en laisses du manuscrit d'Oxford.

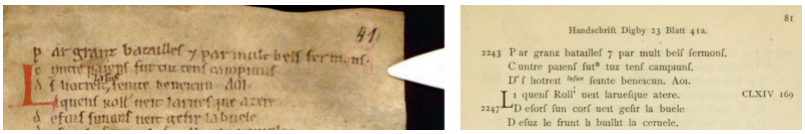


Figure 7: Le haut du feuillet 41 aligné entre le manuscrit et l'ouvrage de Stengel

En utilisant les mécanismes de modification propres aux wikis, nous avons pu transformer l'organisation numérique de l'application [15]. La gestion du manuscrit d'Oxford s'articule maintenant autour d'une première structure hypertexte basée sur les feuillets. A chaque feuillet est associée une page wiki qui est généralement organisée en 3 parties [16] :

— pour le recto, l'association entre le fac-similé de la page du manuscrit et la transcription de Stengel (les liens sur les images sont actifs et permettent des navigations parallèles) ;
— même chose pour le verso ;
— la liste des laisses (avec des liens) avec notre numérotation.

Pour chaque laisse, une page wiki permet de retrouver le ou les feuillets dans lesquels elle est contenue. La suite de l'article montrera qu'elle contient également un ensemble d'informations permettant de confronter les points de vue. Sur cette base, nous allons maintenant aborder deux expérimentations.

— La restitution des annotations de Paul Meyer sur l'édition critique de Francisque Michel. Plutôt destinée à un public de spécialistes (érudits), elle permet de tester l'organisation décrite ici.
— Le traitement de l'oratorio de Gilles Mathieu, et plus particulièrement de son livret afin de permettre au choriste de comprendre le contexte de ce qu'il interprète.

4. Le fonds Paul Meyer pour les spécialistes

En 2014, un étudiant de la filière "Métiers du livre" avait eu pour mission de stage l'exploration et l'analyse de l'édition critique de Francisque Michel de 1869 annotée par Paul Meyer. Suite au travail sur le Palais ducal cité plus haut, le projet Wicri a été sollicité pour aider à produire une version numérique de cette annotation. Ce travail initial a été réalisé, au sein d'un wiki dédié aux collections de la bibliothèque de l'Université de Lorraine, et donc dans un contexte très général.

Cette réalisation est maintenant intégrée dans une bibliothèque spécialisée sur la *Chanson de Roland*, où elle bénéficie d'interactions hypertextes et sémantiques avec les manuscrits et les autres ouvrages sur le sujet.

15. Ceci entraîne naturellement quelques incohérences temporelles mais évite un arrêt de l'application.
16. Exemple le feuillet 41 : https://wicri-demo.istex.fr/Wicri/Europe/ChansonRoland/fr/index.php/Chanson_de_Roland/Manuscrit_d'Oxford/Feuillet_41.

4.1. Le fonds Paul Meyer

La bibliothèque universitaire du Campus Lettres et sciences humaines de l'université de Lorraine à Nancy dispose d'une archive nommée *Fonds Paul Meyer*. Celui-ci, diplômé de l'École des Chartes, philologue et romaniste, spécialiste de littérature romane, a notamment travaillé à la Bibliothèque nationale. Élu au Collège de France en 1876, il prend la direction de l'École des Chartes en 1882. À sa mort, en 1917, il a choisi de léguer sa bibliothèque à l'université de Strasbourg ; mais celle-ci était soumise aux mouvements de frontières que l'Alsace et la Moselle connaissent depuis 1870. C'est donc la bibliothèque de l'université de Nancy qui a été chargée de l'accueillir, par mesure de précaution. C'est ainsi qu'elle abrite le *fonds Paul Meyer*, composé de 4222 titres de monographies et d'environ 7700 brochures, tirés-à-part et petites publications, dont une cinquantaine d'éditions de la *Chanson de Roland*.

4.2. Francisque Michel annoté par Paul Meyer

Dans ce fonds figurent donc plusieurs éditions de la *Chanson de Roland*, dont certaines sont annotées de la main de Paul Meyer.

En 2014, saisissant l'opportunité d'un stage, Isabelle Turcan confiait à l'un de ses étudiants de la filière "Métiers du livre" la tâche d'explorer et d'analyser l'édition de Francisque Michel de 1869 annotée par Paul Meyer. En effet, sur sept pages du recueil, on retrouve des notes, des corrections et des indications d'édition.

4.2.1. Une première expérimentation sur une partie d'ouvrage

Dans cette première expérimentation (en 2014), l'objectif principal était de produire une version OCR correcte d'un texte imprimé avec des annotations.

La mission de stage consistait donc à traiter les annotations manuscrites pour les restituer sur le web. Les pages ainsi traitées présentaient quatre versions :

— un fac-similé de l'original annoté,
— le texte numérisé de Francisque Michel avec ses commentaires,
— le texte avec les annotations de Paul Meyer,
— la version obtenue en intégrant les annotations.

Le travail effectué par l'étudiant ne portant que sur 7 pages, nous avons effectué en parallèle la réédition des 115 autres pages du livre, afin de disposer d'un espace d'expérimentation plus complet.

Dans le contexte ISTEX, deux serveurs d'exploration ont été développés : un sur la *Chanson de Roland*, l'autre sur la philologie.

Enfin, en annotant sémantiquement les variantes des noms de Charlemagne et de Roland, un système d'information a été construit (liste, nombre de pages sur lesquelles chacune est utilisée...) en utilisant des relations sémantiques. Ici 2 types de relations ont été utilisées :

— « `A pour variante de Charlemagne::` » entre une page de F. Michel et la page wiki d'une variable donnée, par exemple « Carles ».
— « `Est une variante orthographique de ::` » entre les différentes variantes et la page Charlemagne.

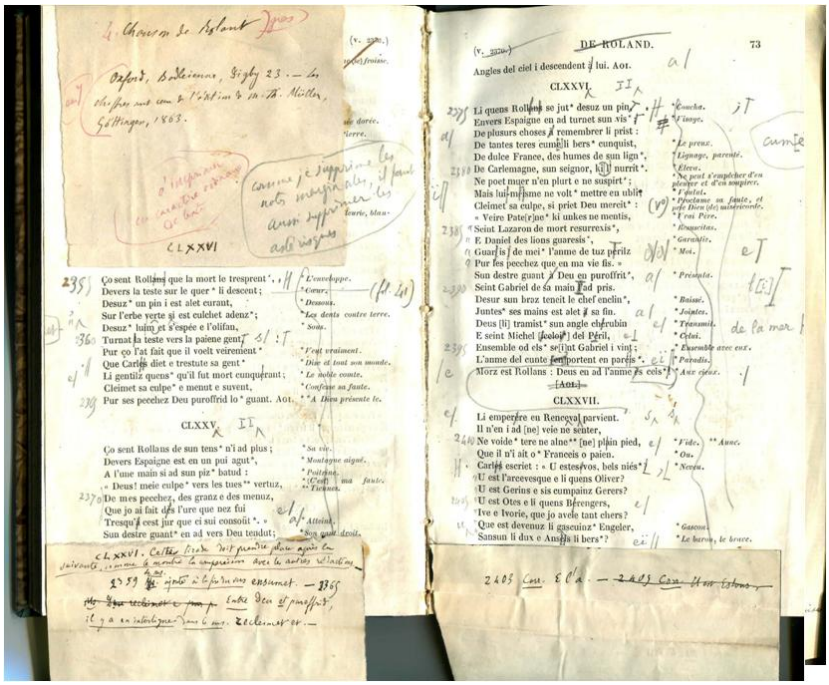


Figure 8: Exemples d'annotations.

Cette première expérience a été montée sur un wiki (collections de la BU Lettres de Lorraine) avec une simple juxtaposition avec d'autres travaux relativement indépendants. Elle est maintenant intégrée à une bibliothèque spécialisée.

4.2.2. Où la bibliothèque ouvre le paysage

Les travaux sur Paul Meyer ont été menés en parallèle avec l'expérience musicale décrite plus loin. Celle-ci repose sur une autre transcription : celle de Léon Gautier.

Nous avons décrit dans la section précédente la gestion des éléments des manuscrits, au départ celui d'Oxford. Les laisses du manuscrit sont devenues un lieu d'interconnexion entre un manuscrit, deux versions critiques et des commentaires.

Sur cette base, nous avons entrepris de compléter le traitement des annotations de Paul Meyer sur Francisque Michel. En effet, l'ouvrage de Francisque Michel contient deux parties. La première est dédiée au manuscrit d'Oxford. La deuxième, nommée *Roman de Roncevaux*, est principalement basée sur le manuscrit de Paris. Elle est également annotée. Nous avons donc décidé de traiter le manuscrit de Paris. Le début de celui-ci est malheureusement tronqué, et Francisque Michel a utilisé le manuscrit de Châteauroux (qui est donc également traité) pour le début de son *Roman de Roncevaux*.

Un chantier est donc en cours pour généraliser l'approche testée avec le manuscrit d'Oxford. Le modèle numérique s'est avéré stable. En revanche la maitrise de l'hétérogénéité des sources est plus complexe. Par exemple, pour le manuscrit de Châteauroux, seule la première page est

disponible avec un fac-similé de bonne qualité à l'IRHT, mais les autres pages, accessibles via le site des bibliothèques de Châteauroux, sont encombrées par une inscription de propriété. En fait chaque manuscrit (Venise, Cambridge) dépend de son propre service de visualisation.

Nous commençons donc à bénéficier d'un dispositif qui permet de confronter deux expertises sur trois manuscrits. L'étape suivante est l'ouverture vers d'autres éditions critiques, et notamment celles de Léon Gautier ou de Joseph Bédier. Les unités numériques de « confrontation » sont naturellement les laisses, mais également les vers et les notes.

Trois principales sources sont actuellement utilisées : Gallica, Internet Archive et Wikisource. Les deux premières offrent un OCR linéaire brut, avec, là encore, des protocoles différents. Wikisource est une source particulièrement intéressante car elle fournit du document « prêt à l'emploi ». Ainsi, avec l'édition critique de Léon Gautier *à l'usage des classes de seconde* publié en 1881, on peut générer un hypertexte de plusieurs milliers de nœuds potentiels (laisses transcrites, traduites, vers, notes sur les vers).

Avec plusieurs documents de ce type, le problème est de concilier une bonne lisibilité par un lecteur humain et la possibilité de réaliser des traitements informatiques. L'approche actuellement testée est basée sur une duplication partielle de ces documents. D'une part, une version arborescente du document est générée en s'appuyant sur les chapitres avec une mise en paragraphe des laisses. Les notes, initialement repoussées dans les annexes (ou dans un autre tome) sont intégrées dans les chapitres numériques. D'autre part, les éléments intéressants sont intégrés dans le graphe des laisses des manuscrits. Ainsi, une laisse dans cet espace expose la diversité des points de vue sans chercher à l'exhaustivité de points de vue communs.

Pour les traitements informatiques, MediaWiki permet d'insérer des annotations en XML. Elles sont notamment utilisées pour réaliser des programmes d'extractions sur des ensembles de pages (sélectionnées par exemple sur un critère sémantique).

4.2.3. Un premier résultat sur les annotations

La valorisation du fonds Paul Meyer a conduit à rechercher ses travaux sur la *Chanson de Roland* pour les intégrer à la bibliothèque numérique. Or Paul Meyer a édité un recueil d'anciens textes bas-latins [2], provençaux et français. On y trouve des extraits relatifs à la *Chanson de Roland*. Nous avons pu constater que des annotations portées sur la version de Francisque Michel se retrouvaient dans le recueil.

4.3. Autour de la revue *Romania*

Paul Meyer est le fondateur de la revue *Romania* qui contient de très nombreux articles sur la *Chanson de Roland*. Ces articles portent naturellement sur l'ensemble des manuscrits et sur les analyses critiques. Ils font de très nombreuses références aux laisses et aux vers. Toutes ces références seront implantées sous forme de liens qui vont compléter cet hypertexte.

Pour le réseau Wicri, cet ensemble devient une description d'un système de recherche dont les relations sémantiques sont relativement classiques. La figure 9 donne un exemple autour d'un article qui traite de « l'accident du vers 2242 ». Ce papier montre comment le copiste a mis par erreur un vers en fin d'un autre feuillet que celui où il devait être copié. La réédition de l'article de *Romania* va donc contenir également des liens vers les manuscrits.

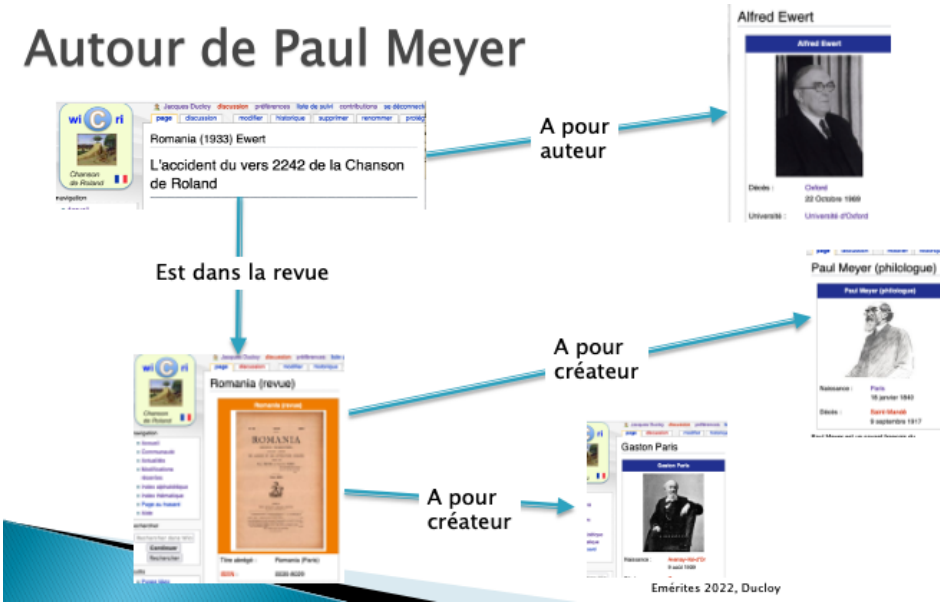


Figure 9: Autour d'un article de la revue *Romania*.

5. Un oratorio pour un public amateur

En complément de ce travail pour philologues, nous souhaitions ouvrir notre réflexion vers un plus large public. Dans un autre contexte, nous avions réédité en hypertexte une messe irlandaise (*Irish Mass*) du compositeur Gilles Mathieu [17]. Or celui-ci a composé un oratorio profane sur la base du manuscrit d'Oxford (dans la version de Léon Gautier). Nous avons donc entrepris d'étudier le rapprochement numérique de la partition et de la transcription du manuscrit.

5.1. Réédition hypertexte d'un oratorio

Pour constituer son oratorio, Gilles Mathieu s'est donc appuyé sur la transcription de Léon Gautier. Il a organisé son livret en dix mouvements. Ceux-ci sont souvent proches de la mise en chapitre de l'ouvrage (exemple : *La cité sur la colline* correspond au *conseil tenu par Marsile à Saragosse*). Il a ensuite sélectionné quelques vers significatifs pour les mettre en musique. Cette musique donne alors un éclairage particulier aux couplets ainsi concernés.

La réédition de l'oratorio va donc contenir des liens vers les laisses correspondantes (avec souvent un décalage de numérotation entre celle qui est citée dans le livret et celle donnée par Wicri). Ainsi, pour chaque mouvement, un paragraphe regroupe, par laisse dans un tableau, l'ensemble des vers utilisés [18].

De plus, l'analyse de la partition montre que, dans un mouvement donné, les phrases musicales

17. https://wicri-demo.istex.fr/Wicri/Musique/fr/index.php/Irish_Mass_(Gilles_Mathieu).
18. Voici un exemple avec le deuxième mouvement : https://wicri-demo.istex.fr/Wicri/Europe/ChansonRoland/fr/index.php/Chanson_de_Roland_(Gilles_Mathieu)/2_-_La_cité_sur_la_colline.

sont généralement associées à une laisse du manuscrit. Pour chaque mouvement, nous avons donc introduit un ensemble de pages de détail, identifiées par un intervalle de mesures. Dans une telle page, les vers sont rappelés avec leur traduction et un pointeur donne accès à laisse correspondante. Les partitions sont données par voix et par instrument, avec une version *tutti*.

Réciproquement, dans chaque laisse concernée, le thème musical est explicité par une ligne mélodique.

Pour la musique, la technologie utilisée repose sur le logiciel de gravure musicale LilyPond. La musique y est codée dans un langage formel dont la syntaxe rappelle celle de TeX pour les mathématiques. Voici par exemple les premières notes du thème « *Au clair de la lune* » en si bémol majeur.

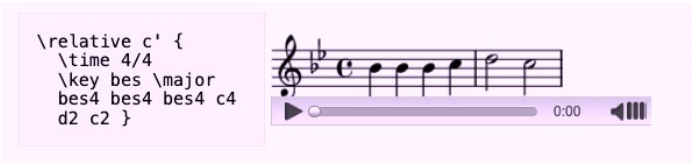


Figure 10: Au clair de la lune en Lilypond.

Ce mode d'interaction permet un travail collaboratif sur une ligne musicale et la réalisation des assemblages en fonction du contexte (présentation d'un thème relatif à un vers du manuscrit ou outil d'apprentissage pour choriste).

Enfin, un blog, installé sur le wiki, et intitulé « dialogue avec un compositeur », permet d'échanger avec Gilles Mathieu sur ses choix musicaux ou sa perception de l'épopée.

5.2. Apports encyclopédiques et sémantiques liés à la vulgarisation

La réédition de cet oratorio veut offrir au choriste, ou au mélomane, une meilleure compréhension du contexte de l'œuvre interprétée ou écoutée. Mais les éditions critiques sont avant tout destinées à un lectorat érudit. Notre bibliothèque doit donc offrir des ouvrages accessibles à un large public.

Le site étant en accès ouvert, les contraintes juridiques limitent très fortement l'utilisation d'éditions modernes [19]. Nous avons réédité une version dite populaire et rédigée par Léon Gautier en 1895 [20]. Elle est effectivement abordable par un public amateur. Mais elle fait appel à de très nombreuses connaissances, parfois décalées (comme les connaissances religieuses entre les XIX^{e} et XX^{e} siècles). Son contenu va donc servir de base pour identifier la base d'un glossaire au niveau du wiki (et pas seulement de l'ouvrage), et ainsi enrichir la base encyclopédique.

Pour améliorer un espace explicatif, un conservateur procède à des acquisitions. Sur Wicri, le documentariste réalise de nouvelles rééditions pour que l'amateur qui découvre le monde des manuscrits puisse en savoir plus. Par exemple, nous envisageons de rééditer le texte d'Eginhard

19. Paradoxalement, les articles de recherche, donc destinés aux érudits, sont plus facilement exploitables avec les nouvelles pratiques de la Science Ouverte.

20. https://wicri-demo.istex.fr/Wicri/Europe/ChansonRoland/fr/index.php/La_Chanson_de_Roland/Léon_Gautier/Édition_populaire/1895.

(*Vita Karoli Magni*) qui cite la bataille de Roncevaux en 830. La même remarque s'applique à Rutebeuf qui cite Roland dans la *Complainte d'Outremer* au XIII^{e} siècle.

De même, l'exploration du paysage correspondant à l'œuvre de Gilles Mathieu conduit à situer cette pièce dans l'histoire poétique et musicale de Roland, comme par exemple l'*Orlando Furioso* de Ludivico Arioso qui a inspiré Vivaldi, Lulli ou Charpentier.

Par rapport à la valorisation du fonds Paul Meyer qui relève d'un contexte professionnel, les besoins de la vulgarisation demandent en fait un approfondissement bien plus important. C'est également vrai sur le plan de la structuration sémantique de la bibliothèque numérique.

Ici, le contexte très spécialisé de *la Chanson* dans le réseau multidisciplinaire Wicri offre un champ d'expérimentation très intéressant. Plus précisément, pour de nombreuses notions, il faut faire cohabiter plusieurs contextes historiques. Par exemple, avec une vision nationaliste, l'Europe au temps de Charlemagne n'est pas celle de notre temps, ni celle de Léon Gautier en 1881 après la Guerre de 1870. La page Europe sur le wiki *Chanson de Roland* sera donc très différente de celle du wiki Wicri/Santé.

Le problème se complique avec les relations sémantiques. Dans pratiquement tous les wikis la capitale de la France est Paris. Ici Paris est bien la capitale de la France pour les auteurs d'articles français. En revanche, l'Empire de Charlemagne, qui n'est pas exactement la France, a pour capitale Aix-la-Chapelle.

Un autre niveau de complexité est introduit par la nature des faits. « Charlemagne, fils de Pépin le bref, est mort à Aix-la-Chapelle » est historiquement vrai. « Roland est mort à Roncevaux » est probablement vrai. « Turpin a été archevêque de Reims » est vrai. Enfin, « Turpin est mort à Roncevaux » est historiquement faux mais légendaire.

6. Bilan et perspectives

Le projet Wicri étudie (au sens preuve de concept) la diversification de l'offre de connaissance scientifique ou culturelle face au monopôle Wikipédia. La figure 11 donne la croissance financière de la Wikimedia Fundation. Elle montre un profond changement de profil financier depuis sa création. Avec un chiffre d'affaires de 120 millions de dollars, basée sur des contributions anonymes, cette compagnie peut-elle garantir le maintien de sa politique citoyenne initiale ?

Sur l'information stratégique, le projet Wicri s'appuie sur les mécanismes d'interopérabilité qui avaient fait le succès des réseaux de coopérations autour des bases Pascal et Francis en France. Ils sont mis en œuvre actuellement pour la santé par le NIH aux États-Unis autour de PubMed. Nous avons un axe de réflexion pour étudier comment un ensemble d'opérateurs européens comme l'INIST pourrait organiser un réseau de valorisation des résultats de la recherche en relation étroite avec les communautés scientifiques.

De façon complémentaire, le projet Wicri/*Chanson de Roland* veut plutôt étudier l'usage des technologies wikis sémantiques (et ingénierie XML) dans les humanités numériques, par exemple pour fédérer les travaux de chercheurs travaillant sur un même sujet.

6.1. Performances techniques

En 2021, les moyens affectés à ces 2 projets ont été limités à un retraité à temps plein et à 2 stagiaires pendant 2 mois (soit 3 semaines de formation pour 15 jours effectifs). Signalons

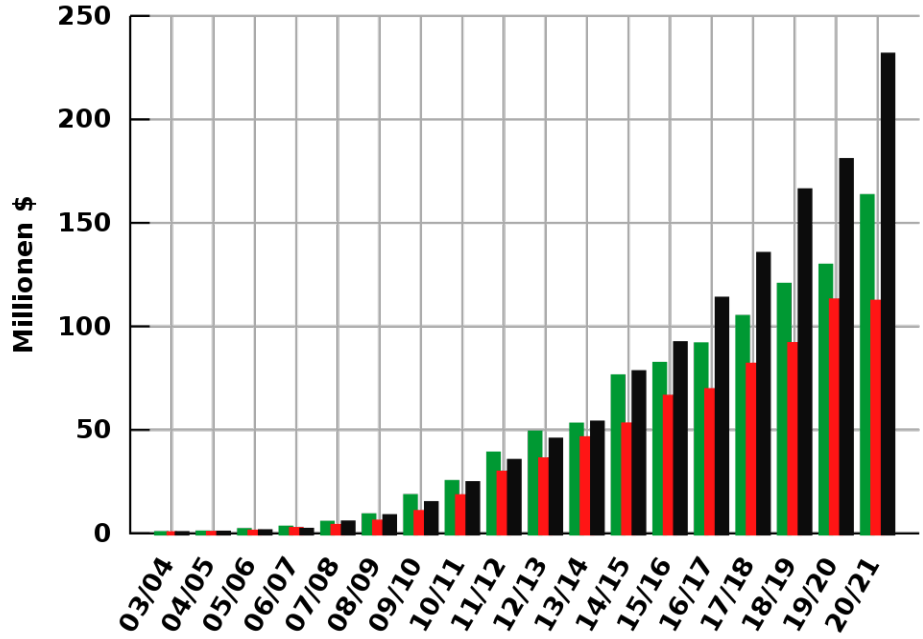


Figure 11: La croissance du chiffre d'affaires de la Wikimedia Fundation.

Table 1
Indices de production sur les wikis (janvier 2022).

	Pages wiki	Avec contenu	Modifications	Sémantique
Chanson de Roland	5 056	1 731	15 738	18 560
Histoire de l'IST	1 839	455	3 798	36 301
Association des émérites	1 562	216	2 250	17 479

également un soutien logistique, limité à quelques demi-journées, mais de haut niveau technique, pour l'hébergement sur le réseau de l'INIST.

Pendant cette période une procédure de changement de version a été entrepris sur une centaine de wikis et trois nouveaux sites significatifs ont été créés (celui sur la *Chanson de Roland*, un sur l'Histoire de l'Information Scientifique et Technique et un autre pour l'association des émérites de Lorraine). Sur ces 3 wikis, le tableau 1 donne les chiffres de production.

Le nombre de pages fait l'objet d'un double comptage. La première colonne donne un nombre total, avec par exemple les pages qui contiennent les modèles ou les déclarations de catégories. Lorsqu'un wiki est initialisé, environ 900 pages de ce type sont chargées (depuis Wicri/Base cité plus haut). La mention « avec contenu » repère les pages de l'espace principal. Cette production a été atteinte en fait avec 6 environ hommes-mois.

Dans le projet Wicri, les mêmes personnes sont donc intervenues sur un ensemble de wikis avec des indices de production significatifs. Les 2 stagiaires (L3 MIASHS [21]) n'avaient jamais travaillé sur des wikis ou dans un environnement Unix. Ils n'avaient aucune connaissance des sujets traités par les wikis. Au bout d'une quinzaines de jours, ils ont pu installer et cataloguer, sur le site des émérites, des dizaines de publications. Au bout d'un mois, ils commençaient à faire des travaux simples mais significatifs sur leurs sujets respectifs (histoire de l'IST en francophonie d'une part, musique et Roland de l'autre), avec un bon début d'autonomie en fin de stage.

Pour la *Chanson de Roland*, les chiffres doivent être rapprochés de la volumétrie des manuscrits. En particulier, l'ensemble des laisses constitue un « plan de travail » qui permet par exemple de rééditer des articles en résolvant les références par des liens hypertextes.

— Le manuscrit d'Oxford contient 170 feuillets, 300 laisses et 4000 vers.
— Un traitement complet de l'ouvrage de Francisque Michel ajoute le manuscrit de Paris (370 laisses) puis le début du manuscrit de Châteauroux (85 laisses).
— L'ensemble des manuscrits de la *Chanson de Roland* se chiffre en centaines de feuillets, en milliers de laisses et en dizaines de milliers de vers.

Le chiffre de 1 731 pages (2 162 en avril) contient notamment la totalité des laisses (et donc des feuillets) du manuscrit d'Oxford. Les annotations de Paul Meyer sont localisées sur les laisses 140 à 160 du manuscrit de Paris. Mais il a été nécessaire de traiter « au sens plan de travail pérenne » les 250 laisses antérieures (plus celles de Châteauroux). Le squelette informationnel concerné par Francisque Michel vient d'être terminé, soit près de 1000 laisses au total. Pour constituer une base pour couvrir un ouvrage hypertexte comparable à celui de Joseph Duggan [11] il faudrait traiter environ 1 000 laisses complémentaires.

La colonne sémantique identifie la production des catégories ou des relations sémantiques. Ceci va être discuté dans la section suivante.

6.2. Relations et web sémantiques

Concernant les aspects sémantiques, les chiffres du tableau montrent une disparité révélatrice. Deux des wikis (IST, émérites) sont relatifs à des systèmes relativement contemporains avec une forte activité éditoriale. L'approche héritée du Semantic Web peut y être déployée. Nous avons cherché à la compléter à partir des approches CRIS [22] et plus précisément du modèle CERIF [12]. Dans les deux cas nous avons rencontré quelques difficultés liées à la francisation de ces systèmes. De même, la généralisation de ce modèle à toutes les disciplines scientifiques demande des adaptations (ou des simplifications) notamment dans les humanités qui ont une grande variété de fonctionnement de comités scientifiques ou éditoriaux.

Dans le cas de la *Chanson de Roland*, le modèle sémantique de ce sujet médiéval, analysé pendant plusieurs siècles, doit être construit quasi intégralement. Pour cela, l'utilisateur, en situation de concepteur, dispose de nombreux outils qu'il peut combiner :

— MediaWiki offre un mécanisme d'indexations hiérarchisées à base de catégories.
— Celles-ci peuvent être crées ou manipulées avec des modèles ou des modules (en langage informatique Lua).

21. Licence en mathématiques et informatique appliquées aux sciences humaines et sociales.
22. *Current Research Information Systems* (systèmes d'informations sur les recherches en cours).

— Semantic MediaWiki permet de créer des triplets sémantiques. Ils peuvent également être manipulés avec les modèles et combinés avec les catégories.
— Dilib, la boîte à outil XML initialement conçue pour des analyses statistiques de corpus qui contient maintenant des modules d'interface qui peuvent utiliser l'API de MediaWiki.

Cet ensemble est naturellement utilisable de façon incrémentale.

Le modèle sémantique de la *Chanson de Roland* repose déjà, pour chaque manuscrit, sur un graphe fiabilisé au niveau des feuillets et des laisses. Ce travail implique une numérotation fiabilisée, et donc un traitement séquentiel de l'ensemble manuscrit (évoqué plus haut). Le résultat est un ensemble de pages identifiées par une nomenclature arborescente, par exemple :

`Chanson de Roland/Manuscrit d'Oxford/Laisse CCX`

Cette nomenclature permet déjà de piloter un robot. Un chantier est maintenant ouvert sur l'élaboration d'un ensemble sémantique permettant à un contributeur de poser des requêtes au sein du wiki. Les modèles permettent déjà de créer de façon implicite des relations normalisées. Par exemple, la figure 12 montre un bandeau qui est généré par l'appel suivant :

```
{{Manuscrit de Paris/Header laisse | sort=004 | id=IV | précédent=III |
                                suivant=V}}
```

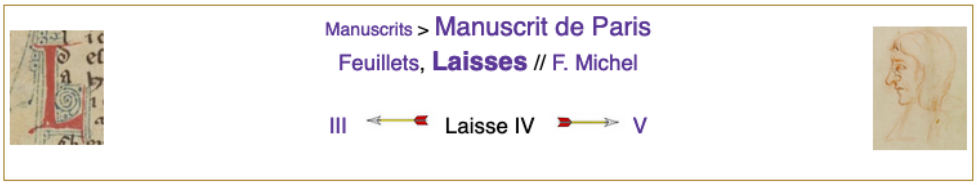


Figure 12: En tête généré par l'appel d'un modèle.

Cet appel va provoquer l'ajout d'une catégorie « Laisses » et une relation sémantique « `A pour manuscrit::` » avec la page « `Chanson de Roland/Manuscrit de Paris` ».

Cette organisation sémantique doit être maintenant confrontée avec les applications comme par exemple l'étude des annotations de Paul Meyer. Puis il faudra identifier les relations génériques ou celles qui doivent être diversifiées. L'apport fondamental de l'approche wiki est la possibilité d'une démarche incrémentale où le modèle peut être élaboré dans un processus « essai – erreur ».

Par rapport au Web sémantique, nous avons déjà mentionné des stratégies d'alignement sur des terminologies existantes, notamment Wikipédia en français qui est aligné avec son équivalent anglais souvent présenté comme base terminologique du Web sémantique.

Cela dit, nous avons pu constater la difficulté d'alignement sur des sites de référence comme Worldcat de l'OCLC. Nous avons notamment essayé d'identifier les versions de Léon Gautier « de son vivant » soit avant 1897. Une requête, pas forcément exhaustive, donne 82 entrées sur 129 au total [23]. La version 1881 contient dans son titre la mention « 23e édition ». Il y a donc

23. https://www.worldcat.org/search?q=chanson+de+roland+léon+gautier.

environ 60 doublons sur un ensemble de 80 notices. Plus inquiétant, 30 entrées ont 1872 comme date d'édition. Or la revue *Romania* donne en 1873 une analyse relative à la troisième édition datée de 1872 (Paris 1873) et qui, de plus, n'a pas été imprimée... Donc, 30 notices portent en fait sur 2 exemplaires réels. Ces exemples montrent les limites d'une utilisation non contrôlées des triplestores produits par accumulation de sources de données.

En revanche, pour le réseau des communautés du réseau Wicri, les stratégies collaboratives basées sur des noyaux de métadonnées « modérables » semblent plus prometteuses. Signalons par exemple les travaux des Smithsonian Institutions [13] aux États-Unis, du fichier d'autorité commun (GHD) de la Bibliothèque Nationale d'Allemagne [14], et de la Bibliothèque nationale de France [15]. Les trois approches sont basées sur des solutions MediaWiki avec l'extension WikiBase. Dans ces trois cas, il s'agit de mutualiser des catalogues de collections. Le projet Wicri s'attaque à un autre problème : la mutualisation de connaissances, portées par des articles présélectionnés. Le modèle développé pour des collections, grandes mais finies, doit être revisité pour des connaissances potentiellement infinies.

Un article de Luca Mauri [16] étudie (favorablement) la cohabitation des extensions Semantic MediaWiki et WikiBase. Il met également en avant, pour Semantic MediaWiki, la possibilité offerte au contributeur de formuler des requêtes.

A l'heure actuelle, pour le projet Wicri, l'ensemble des wikis est implanté sur la même machine virtuelle (et donc sur la même machine physique). Il parait donc souvent plus efficace d'utiliser des procédures batch. Dans une évolution vers un réseau physiquement distribué, la situation sera naturellement différente, et l'ouverture apportée par WikiBase semble effectivement séduisante.

6.2.1. Modèles sémantiques propres à l'histoire, à la culture et à la légende

Nous venons d'évoquer principalement des relations relevant plutôt de la modélisation de systèmes de recherche (ou éditoriaux). Nous y avons appliqué des outils et pratiques élaborées dans d'autres disciplines scientifiques notamment dans la santé ou de l'environnement. Or la *Chanson de Roland* est un sujet majeur dans les langues romanes et autour de l'histoire ou des légendes de Charlemagne. Elle est donc un sujet privilégié pour l'étude des approches sémantiques sur les données textuelles des humanités numériques.

Dans un premier temps, nous avons traité des articles scientifiques de référence (de la revue *Romania* par exemple). Nous avons ainsi consolidé les connaissances autour du réseau d'acteurs spécialistes du sujet et montré l'intérêt des liens vers les laisses des manuscrits. Nous avons identifié d'autres documents ou manuscrits à ajouter à notre bibliothèque numérique. Nous avons par exemple commencé à traiter un manuscrit en allemand : le *Rolandslied* du curé Konrad, écrit vers 1170. Malheureusement, nos compétences étaient encore trop approximatives pour réaliser un travail d'indexation pertinent, compte tenu du niveau d'érudition du contenu des articles.

En revanche, comme évoqué plus haut, nous avons démarré un travail qui apparait très prometteur avec la réédition numérique de l'édition populaire publiée par Léon Gautier en 1895. En effet, cet ouvrage contient un très grand nombre de notes explicatives qui identifient un ensemble de concepts essentiels. Les sujets sont multiples : personnages, lieux géographiques, particularités linguistiques, visions historiques.

Nous démarrons donc des expérimentations avec les personnages. Faciles à identifier, ils font émerger immédiatement les difficultés (et l'intérêt) de la multiplicité des situations, des types de discours et des besoins des contributeurs.

Voici quelques exemples sur Roland et Charlemagne. Ces deux héros sont omniprésents sur ce wiki et seules les relations spécifiques sont *a priori* intéressantes. Il faut donc distinguer par exemple les textes narratifs et les articles historiques. Ensuite, les textes narratifs sont de style et de granularité différentes. Les personnages peuvent y apparaitre sous leur référence historique ou avec un éclairage légendaire. Nous avons introduit deux relations sémantique différentes pour marquer cette différence, notamment dans les pages affectées aux laisses des manuscrit.

Ainsi la première laisse du manuscrit d'Oxford contient « Charles le roi ... sept ans est resté en Espagne ». Sa présence en Espagne est historiquement juste, même si la durée de sa campagne est plus modeste. Nous avons donc choisi d'affecter à cette laisse un attribut préexistant :

Laisse 1 > A pour personnalité citée : : Charlemagne

En revanche, dans la laisse CIX, le vers 1404 nous dit « Charles le Grand en pleure et se lamente ». Nous sommes ici dans la légende et nous avons introduit un nouvel attribut indiquant l'aspect « personnage de légende ». L'attribut devient :

Laisse CIX > A pour personnage cité : : Charlemagne

Dans certaines parties du récit, les personnages parlent. Ils peuvent même chanter dans l'oratorio. Nous trouverons alors des attributs tels que :

```
Mouvement 3 /Mesure 29 à 35 > A pour personnage chantant:: Roland
```

La présence de cet attribut est limitée aux pages dédiées aux courtes séquences musicales (et non, par exemple, aux descriptions des mouvements.

Comme indiqué plus haut, ce travail est en phase de démarrage. En effet, pour être pertinent il doit s'appuyer sur l'ensemble des laisses d'un manuscrit donné. Nos premières observations montrent que cette indexation va être très dépendante des préoccupations scientifiques des utilisateurs de cette bibliothèque. Plusieurs modèles sémantiques devront probablement cohabiter.

Par rapport aux travaux sur le web sémantique, une première réflexion semble s'imposer. Il est relativement facile d'appuyer sur un bouton de paramétrage pour offrir dès maintenant 25 000 triplets RDF. Mais, de notre point de vue, ceci n'aurait aucun sens. Par rapport aux offres de contenus relativement homogènes (exemple un site dédié aux partitions musicales) notre bibliothèque propose des ensembles de ressources différentes sur un même sujet. Il faudra donc plutôt imaginer un ensemble de triplestores.

6.3. Aspects multimédias

Le traitement des illustrations actuellement utilisé dans le réseau Wicri est relativement classique : des insertions d'images, avec parfois, notamment pour les serveurs d'exploration, des cartes interactives. Pour naviguer dans les bases d'images, il y a une vingtaine d'années,

nous avions utilisé Dilib pour générer des graphes de navigation [17]. Sur la plupart des wikis, ces images sont des objets relativement indépendants les uns des autres. Nous n'avons donc pas été incités à faire de la navigation dans les images. En revanche, sur Wicri/*Chanson de Roland*, les éditions populaires offrent une variété d'images sur un même thème avec, par exemple des graveurs de référence comme Merson, Ferat ou Zier. Un projet dans cette direction pourrait rapidement être mis en place.

Concernant la musique, sur Wicri/Musique, nous avons repris les extensions basées sur LilyPond avec un double point de vue : permettre à un lecteur de se faire une idée d'une pièce musicale et offrir à un choriste des outils de répétition. Avec la *Chanson de Roland*, les extensions musicales sont aussi utilisées pour accompagner le texte des laisses ; ce qui est bien perçu lors des démonstrations. Dans l'avenir, un point fondamental doit être abordé : un éclairage sur la prononciation des textes en ancien français. Or de nombreux articles traitent de ce sujet autour de la *Chanson de Roland*. La liaison entre les aspects phonétiques de la langue des manuscrits avec la musique chantée ouvre un champ d'application très intéressant [24].

La notation LilyPond offre également une possibilité que nous n'avons pas encore utilisée : la recherche de séquences ou de particularités musicales.

Dans ces différents exemples, la précision de l'indexation et des modèles sémantiques joue un rôle fondamental.

6.4. Gérer l'incomplétude et la multiplicité des besoins

L'approche wiki permet de diffuser très rapidement des premiers résultats, même inachevés. L'intérêt très clair : des non-spécialistes de la technologie bénéficient ainsi d'un substrat concret sur lequel ils peuvent immédiatement travailler. Ainsi, un expert de la musicologie, un linguiste, ou un médiéviste, peuvent rapidement s'emparer du projet sans avoir à passer par la technique. Le revers de cette médaille est la gestion de l'incomplétude qui devient un problème omniprésent.

Dans le wiki sur la *Chanson de Roland*, la volumétrie est déjà consistante. Une amélioration minime sur le contenu des laisses (qui demanderait par exemple 2 minutes par action) peut se traduire par des dizaines d'heures de travail. Cela dit, Wikipédia rencontre des problèmes analogues et sait les traiter en organisant des chantiers (ou en programmant des robots). Le même type d'approche doit pouvoir se dégager ici.

Deux types de chantiers, sont amenées à coexister. Pour certaines opérations, comme cité plus haut « numéroter les laisses d'un nouveau manuscrit », il est indispensable de travailler dans une continuité totale. À l'inverse, certaines expérimentations ont, par nature, une nature transversale, et nécessitent de parcourir quelques pages sur lesquelles sont effectuées des opérations ponctuelles. L'enjeu majeur devient alors d'assurer la cohérence du traitement malgré son émiettement.

Dans les deux cas, les visiteurs peuvent être confrontés, lors d'une exploration inopinée, à des erreurs, à des liens brisés, à une navigation rendue complexe par des situations de « rupture de phase ». Il faut donc travailler sur la constitution de complétude partielle autour de thèmes démonstratifs. Cela souligne la forte dimension éditoriale, qui ne peut et ne doit pas être négligée.

24. Voir un premier exemple : https://wicri-demo.istex.fr/Wicri/Europe/ChansonRoland/fr/index.php/La_Chanson_de_Roland/Léon_Gautier/Édition_populaire/1895/Introduction/La_versification.

Enfin, les moyens affectés au projet Wicri sont relativement insignifiants par rapport aux enjeux. Nous donnons donc la priorité aux aspects « preuve de concepts » dans la variété des disciplines scientifiques. Nous assumons une situation où nous laissons un sujet en sommeil quand il ne pose plus de difficultés, pour aborder un autre problème qui nous parait important dans une stratégie de déploiement institutionnel.

6.5. Aspects institutionnels, du contrôle à l'accompagnement

Le premier auteur de cet article a exercé des responsabilités à l'INIST (informatique, R&D, produit et services) et dans un centre de calcul partenaire du Trésor de la langue Française dans les années 75 [18]. Ces deux unités du CNRS ont depuis renoncé à leurs missions citoyennes (notamment pour le TLF) ou stratégiques (Pascal) après avoir été confrontées à des difficultés de production. Une motivation fondamentale du projet Wicri est donc la recherche de solutions qui permettraient de reprendre ce type d'activités avec une vision européenne.

Dans les deux cas (INIST ou TLF), la simple gestion des chaines de production demandait des forces de développement et d'exploitation considérables (par exemple, à l'INIST en 2000, la taille du département informatique a dépassé 50 personnes). Notre expérience, comme celle de Wikipédia, montre que ces investissements peuvent être réduits de façon considérable. En revanche, cette approche demande un haut niveau d'expertise informatique (par exemple la capacité de réaliser un robot dans une structure arborescente de documents structurés).

Dans ces deux cas, les procédures de contrôle alourdissaient, de façon parfois ahurissante, la productivité (et la motivation) des ingénieurs. Par exemple, à l'INIST, les protocoles définis par une cellule de qualité étaient conçus pour des applications financières. La simple correction d'une faute d'orthographe dans un résumé monopolisait deux techniciens avec contrôle de leurs hiérarchies (autrement dit, des dizaines de minutes d'intervention et des mois de délai pour une opération qui ne demande que quelques secondes sur un wiki).

Concernant les activités scientifiques proprement dites, les tentatives de remplacement des ingénieurs par des algorithmes se sont traduites par des échecs qui ont abouti à l'arrêt de Pascal (là où le NIH [25] aux USA soutient la production par la *National Libray of Medicine* de 900 000 analyses documentaires par an où chaque article est lu par 3 experts avec une indexation assistée — mais pas automatique). La technologie wiki permet de faire travailler un réseau d'experts sur un ensemble d'applications (élaboration d'ontologies, édition encyclopédique, indexation, secrétariat de rédaction) dans le même environnement et en mode réseau. Elle demande une expertise technique et multidisciplinaire plus élevée que celle des chaînes de production. Notre expérience (par exemple avec les stagiaires) montre que les protocoles d'assistance et de modération fonctionnent, à condition qu'ils soient continus. Un dispositif de taille comparable à celui de Pascal/Francis en 1990 (400 personnes) est donc nécessaire, mais avec un mode de fonctionnement totalement différent.

Un tel dispositif doit être le plus proche possible des centres de recherche. Pour cela, la production de revues permet de créer des comités scientifiques qui peuvent être mobilisables dans la modération du wiki (nous avions testé cette approche avec la revue *AMETIST*). Dans les activités encyclopédiques, il est également envisageable d'impliquer des thésards dans leur phase d'étude de l'existant (avec la participation de leurs directeurs de thèse).

25. NIH : National Institutes of Health, institutions qui dépendent du Département de la Santé aux USA.

De son côté, l'expérimentation autour de la *Chanson de Roland* montre une exigence d'expertise multidisciplinaire. Là encore, un dispositif d'accompagnement conséquent, notamment en termes de permanence, s'avère indispensable. Il n'est pas à la portée d'une équipe de recherche ou d'un petit laboratoire. En revanche, il peut être obtenu par mutualisation d'équipes au sein d'une université, autour d'une Maison des sciences de l'homme, ou avec une bibliothèque universitaire en mode *learning center*.

Bien entendu, une organisation européenne (et/ou francophone), avec une solution telle que celle qui est étudiée avec Wicri, peut jouer un rôle d'accompagnement pour des projets, même de petite taille dans un contexte international. Par exemple, si la reprise en ordre de marche d'initiatives porteuses des ambitions de Pascal (informer sur l'essentiel de la science) peut être imaginée par discipline scientifique, avec si possible, une mutualisation des expertises.

7. Conclusion

Nous avons présenté un projet numérique autour de la *Chanson de Roland* dans une infrastructure multidisciplinaire.

Notre projet montre d'abord l'intérêt de nouvelles approches, sémantiques, hypertextuelles, pour les bibliothèques — et les bibliothèques numériques — dans le contexte des humanités numériques. Il met également en évidence l'explosion de nouvelles barrières dans un changement de paradigmes dans les mondes de la recherche ou de la connaissance.

La *Chanson de Roland* nous fait voyager dans le temps entre le Moyen-Âge et le troisième millénaire en passant par le XIX^{e} siècle.

Au temps de Roland, les lettrés, copistes ou bibliothécaires, disposaient d'une très grande marge d'initiative qu'ils ont perdu parfois avec l'imprimerie. Le numérique leur permet aux humanistes de retrouver cette autonomie. En particulier, les rééditions hypertextuelles revisitent les pratiques des copistes qui avaient une part d'interprétation d'un texte à la façon d'un musicien sur une partition.

Au temps de Roland, les sciences de la matière étaient dominées par les alchimistes qui sont devenus chimistes en s'appropriant des outils mathématiques de plus en plus sophistiqués, au XIX^{e} siècle avec les équations différentielles, et maintenant avec le numérique des big data. Avec le numérique les chercheurs et praticiens des humanités doivent se dégager de la domination des informaticiens en s'appropriant à la fois, l'algorithmique, les techniques sémantiques et les manipulations avancées de corpus...

Tout un programme qui peut s'avérer passionnant !

Remerciements

Nous remercions vivement Gilles Mathieu pour sa coopération constante sur le projet. Merci aux valeureux stagiaires Dalila Ladli et Léonard Braux qui ont défriché le terrain. Merci aux équipes techniques de l'INIST, à sa direction et aux instances d'ISTEX pour l'hébergement du réseau Wicri. Merci aux groupes de travail Wicri pour leur soutien amical.

Références

[1] F. Michel, La chanson de Roland ou de Roncevaux du XIIe siècle, Silvestre, Paris, 1837.

[2] P. Meyer, Recueil d'anciens textes bas-latins, provençaux et français, F. Vieweg, Paris, 1874.

[3] L. Gautier, La chanson de Roland, Traduction, précédée d'une introduction et accompagnée d'un commentaire, par Léon Gautier, Édition populaire, Tours, 1895.

[4] J. Ducloy, T. Daunois, J.-P. Thomesse, F. Peguiron, I. Turcan, Revisiter les textes anciens dans les bibliothèques numériques avec l'exemple de la chanson de roland, 2021. URL : https://wicri-demo.istex.fr/Wicri/Europe/ChansonRoland/fr/index.php/HIS_2021_Casablanca/Atelier_Wicri/Article_soumis.

[5] S. Vahdati, N. Arndt, S. Auer, C. Lange, Openresearch : Collaborative management of scholarly communication metadata, in : E. Blomqvist, P. Ciancarini, F. Poggi, F. Vitali (Eds.), Knowledge Engineering and Knowledge Management, Springer International Publishing, Cham, 2016, pp. 778–793.

[6] J. Ducloy, Lorexplor : une bibliothèque open source de composants xml d'exploitation du corpus. séminaire de bilan du projet istex, 2018. URL : https://wicri-demo.istex.fr/Wicri/Wicri/fr/index.php/Utilisateur:Jacques_Ducloy/Blog/Séminaire_ISTEX_2018.

[7] J. Ducloy, P. Charpentier, L. Grivel, Une boîte à outils pour le traitement de l'information scientifique et technique, Génie logiciel et systèmes experts 25 (1991) 80–90.

[8] J. Ducloy, T. Daunois, M. Foulonneau, A. Hermann, J.-C. Lamirel, S. Sire, J.-P. Thomesse, C. Vanoirbeek, Metadata for wicri, a network of semantic wikis for communities in research and innovation, in : DC-2010–Pittsburgh Proceedings, 2010, pp. 94–102.

[9] J. Horrent, La chanson de Roland dans les littératures française et espagnole au Moyen Âge, Presses universitaires de Liège, Liège, 1951, pp. 323–333. URL : https://books.openedition.org/pulg/1327.

[10] E. Stengel, Das altfranzösische Rolandslied. Genauer Abdruck der Oxforder Hs. Digby 23, Gebr. Henninger, Heilbron, 1878.

[11] J. Duggan (Ed.), La Chanson de Roland. The Song of Roland. The French Corpus, Brepols, Turnhout, 2006.

[12] O. Azeroual, J. Schöpfel, Quality issues of cris data : An exploratory investigation with universities from twelve countries, Publications 7 (2019). URL : https://www.mdpi.com/2304-6775/7/1/14. doi :10.3390/publications7010014.

[13] J. Shieh, Smithsonian libraries and archives & wikidata : Using linked open data to connect smithsonian information, 2022. URL : https://blog.library.si.edu/blog/2022/01/19/smithsonian-libraries,Archives&Wikidata:UsingLinkedOpenData/archives-wikidata-using-linked-open-data-to-connect-smithsonian-information/#.YfKzg2DjKlM.

[14] B. Fischer, Towards an open and collaborative authority control, JLIS.it 13 (2022) 283–290. URL : https://jlis.it/index.php/jlis/article/view/438. doi :10.4403/jlis.it-12767.

[15] V. Boulet, How to build an «identifiers' policy» : the bnf use case, JLIS.it 13 (2022) 177–184. URL : https://www.jlis.it/index.php/jlis/article/view/429. doi :10.4403/jlis.it-12768.

[16] L. Mauri, The best of both worlds : Wikibase and semantic mediawiki, 2021. URL : https://mediawikiexperts.blog/the-best-of-both-worlds-wikibase-and-semantic-mediawiki/.

[17] J.-C. Lamirel, J. Ducloy, G. Oster, Adaptive browsing for information discovery in an

iconographic context, in : Content-Based Multimedia Information Access - Volume 2, RIAO '00, LE CENTRE DE HAUTES ETUDES INTERNATIONALES D'INFORMATIQUE DOCUMENTAIRE, Paris, FRA, 2000, p. 1657–1675.

[18] J. Ducloy, Systèmes d'information encyclopédiques édités par les scientifiques, Revue ouverte d'ingénierie des systèmes d'information 1 (2020). URL : http://www.openscience.fr/Systemes-d-information-encyclopediques-edites-par-les-scientifiques. doi :10.21494/ISTE.OP.2020.0488.

Bibliographie philosophique et humanités numériques : de la cartographie des sciences à l'encyclopédie opérationnelle

Philosophical bibliography and digital humanities: from the mapping of sciences to the operational encyclopaedia

Fabien Ferri[1,*], Tom Annebi[1]

[1]*Université de Franche-Comté, Laboratoire Logiques de l'Agir, F-25000 Besançon, France*

Abstract

This article proposes to study the evolution of a bibliographic medium, the *Bibliographie de la philosophie*, as it has had to transform itself, following the impact of digitisation on the material bibliography resulting from the print culture. We present the way in which this transformation was enacted within a French bibliographic processing centre, the *Centre de documentation et de bibliographie philosophiques* of the University of Franche-Comté (CDBP), through the implementation of a bibliographic database in philosophy of science : the *Système d'information en philosophie des sciences* (SIPS). We thus aim to describe two processes : 1° the refoundation of a bibliographic enterprise as it was impacted by the technological shock of digitisation ; 2° the way in which this enterprise was transformed to provide a support tool for research in the digital humanities. We sketch a perspective of the evolution of the system in which the current cartography of sciences could interface with an operational visual encyclopaedia.

Keywords

bibliography, encyclopaedia, digital humanities, philosophy, documentary information system

1. Introduction : informatisation, numérisation et humanités numériques

Nous parlons en français d'humanités numériques. Dans ce syntagme, le terme numérique provient du latin *numerus*, qui désigne le nombre au sens arithmétique, c'est-à-dire tel qu'il renvoie à la cardinalité d'un ensemble fini d'éléments. Un ensemble fini d'une même classe d'êtres est par définition numérique [1]. Les anglais parlent quant à eux de *digital humanities* [2]. Digital provient de *digit*, qui veut dire « le doigt ». Dans le premier cas, *numerus*, on associe le nombre à la cardinalité d'un ensemble abstrait. Dans le second cas, *digit*, on indexe le nombre au sens très fort du terme, puisqu'on fait correspondre l'unité arithmétique élémentaire à l'index de la main, c'est-à-dire à une entité concrète directement manipulable, puisque constitutive de la main : un doigt.

Mais le terme « numérique » dans le syntagme « humanités numériques » ne renvoie ni à la cardinalité d'un ensemble abstrait, ni à l' « indexicalité » d'un élément concret. Il renvoie à un

. *Workshop on Digital Humanities and Semantic Web*

*. Corresponding author.

. fabien.ferri@univ-fcomte.fr (F. Ferri) ; tom.annebi@edu.univ-fcomte.fr (T. Annebi)

. https://logiquesagir.univ-fcomte.fr/fabien-ferri/ (F. Ferri)

support de codage discret, opératoire et calculatoire [3]. Les humanités numériques, ce sont donc les lettres et belles-lettres codées par un système de nombre qu'on appelle le binaire. Ce sont aussi les sciences humaines dans toute leur diversité mais aussi la philosophie, qui, une fois codées par le système binaire, sont soumises à des processus calculatoires. Les résultats des calculs exécutés par les machines computationnelles sur les codages des contenus des humanités *imprimées* et *analogiques*, issus d'une part des supports technologiques de la culture de l'écrit [4] et d'autre part des supports dérivés des technologies culturelles d'enregistrement analogiques [5], correspondent à des contenus documentaires numérisés (son, image, texte, image animée, etc.), c'est-à-dire régénérés sur des interfaces virtuelles et des terminaux de sorties [6]. Leurs formes sémiotiques sont perceptibles et appropriables soit à travers des écrans, soit à travers des haut-parleurs. Dans cette mesure, le numérique constitue un système technique homogène [7, p. 15] et un support universel de codage de l'information destiné à produire des formes sémiotiques perceptivement appropriables par un public [8].

Cependant, le numérique ne doit pas être confondu avec l'informatique. Il en dérive, mais ne s'y réduit pas. L'informatique, comme le mot l'indique dans sa contraction [9, 10], désigne en effet le traitement auto*matique* de l'*infor*mation comprise au sens technique du terme, c'est-à-dire tel que défini par Claude Shannon au sortir de la Seconde Guerre mondiale [11, p. 16]. L'information, en ce sens technique, c'est un code. Or un code est d'autant plus informant qu'il contient de bits. C'est pourquoi l'informatique a pu être définie comme une physique des signes sans significations (les bits) et abstraite (car sans substrat) mais aussi comme la technologie des calculs effectuables sur des codages, c'est-à-dire comme traitement automatique de l'information numérisée [11, p. 16].

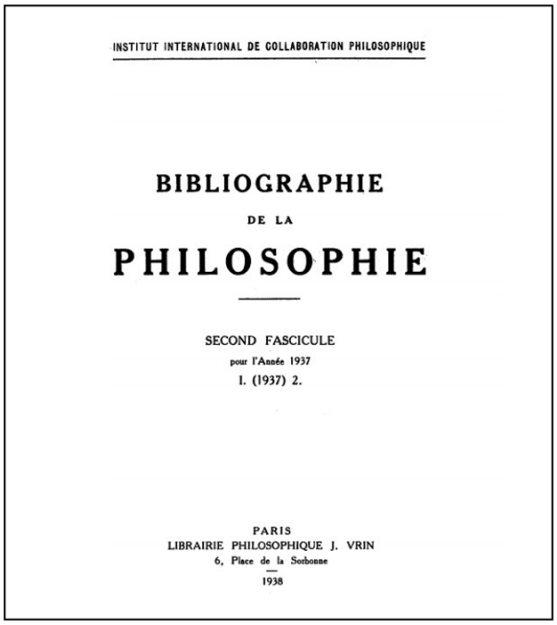
La question que nous posons consiste alors à savoir comment cette technologie a impacté la culture traditionnelle de l'écrit – et donc les humanités traditionnelles – depuis l'invention de l'imprimerie, et cela en raison : 1° de la possibilité de la numérisation des contenus ouverte par le codage binaire de l'information ; 2° de la possibilité de calculer ces contenus dès lors qu'ils sont codés sur ce support virtuel opératoire et calculatoire qu'est la machine de Turing [12].

2. *Bibliographie de la philosophie* et bulletin bibliographique imprimé

Fondé en 1937 à l'initiative du philosophe suédois Åke Petzäll grâce au soutien des philosophes Raymond Bayer, Émile Bréhier et Léon Robin, mais aussi grâce à une collaboration institutionnelle entre la Sorbonne et l'Université de Lund, l'*Institut international de philosophie* est actuellement une association philosophique cosmopolite composée d'une centaine de membres [1]. Raymond Klibansky était jusqu'en 2005 le président de la Commission des travaux bibliographiques de cette association [13], ainsi que le responsable de la publication trimestrielle internationale dont elle rendait possible la coordination à l'échelle planétaire : la *Bibliographie de la philosophie* (Figure 1). Créé en 1959 par Gaston Berger et Gilbert Varet, le *Centre de documentation et de bibliographie philosophiques* de l'université de Besançon (CDBP) fut l'un des deux centres de traitement en France consacrés à la mise en œuvre de ce projet éditorial,

1. https://www.i-i-p.org/

dont le but était de constituer une bibliographie analytique internationale dans le champ de la philosophie, ce qui fut le cas entre 1937 et 2010 [2].


INSTITUT INTERNATIONAL DE COLLABORATION PHILOSOPHIQUE

BIBLIOGRAPHIE

DE LA

PHILOSOPHIE

SECOND FASCICULE

pour l'Année 1937

I. (1937) 2.

PARIS

LIBRAIRIE PHILOSOPHIQUE J. VRIN

6, Place de la Sorbonne

1938

Figure 1 : Second fascicule de la *Bibliographie de la philosophie*, 1938.

Parallèlement à cette entreprise, le *Philosopher's Index* lancé en 1967 par le *Philosopher's Information Center* était centré sur la production anglo-saxonne [3]. Ainsi, face à un premier projet, la *Bibliographie de la philosophie*, qui n'avait pas acté l'impact de l'informatisation et de la numérisation sur l'activité bibliographique et à un second, le *Philosopher's Index*, qui ignorait la période décisive de la philosophie des sciences (la première moitié du 20e siècle car témoin de la naissance de ce domaine de la recherche philosophique), il s'agissait de réorienter l'activité du traitement bibliographique. D'une part vers des supports informatiques et numériques ; d'autre part en appliquant ce traitement à un domaine non répertorié de façon systématique depuis le 19e siècle, ce qui était le cas de la philosophie des sciences.

3. L'arrêt de la *Bibliographie de la philosophie* et la naissance du SIPS

Afin de procéder à l'identification du domaine de la philosophie des sciences, l'idée au fondement du *Système d'information en philosophie des sciences*, qu'on doit à Thierry Martin, fut de constituer une base de connaissances couvrant la totalité des publications en philosophie des sciences pour les 20e et 21e siècles, pour en dresser le recensement analytique. Un tel système

2. https://www.i-i-p.org/bibl
3. https://philindex.org/about-us/history/

n'existe toujours pas aujourd'hui, que ce soit à un niveau national, européen ou international [4].

À partir de 2010, il s'agissait ainsi pour le projet SIPS de commencer à se positionner en qualité de fournisseur de données, grâce à la mise à disposition d'un ensemble de notices bibliographiques dans un catalogue XML exposant des métadonnées définies en Dublin Core simplifié, conformément au protocole OAI-PMH et en accord avec les recommandations du Très Grand Équipement Adonis [14]. En devenant un entrepôt, le SIPS pouvait s'assurer une visibilité optimale auprès des fournisseurs de services tels que Gallica [5], HAL [6] ou Isidore [7], pour que ces derniers viennent moissonner son serveur de données afin d'inclure ses notices bibliographiques numériques dans leurs propres résultats de recherche.

La construction de l'application informatique supportant la base de données bibliographiques et la réalisation des interfaces administrateurs et utilisateurs a dès lors été réalisée grâce à une collaboration entre le CDBP et plusieurs informaticiens. Plus de 2000 notices bibliographiques avaient alors déjà été rédigées au sein du CDBP. Parallèlement à ces activités techniques, l'organisation d'un séminaire mensuel de philosophie des sciences dans le cadre des activités de l'équipe de philosophie de l'université de Franche-Comté [8], la tenue du second congrès de la Société de philosophie des sciences à Genève du 29 au 31 mars 2007, qui portait sur le thème de la question de l'unité des sciences aujourd'hui [9], et enfin l'organisation à Besançon de journées d'études sur la scientificité des sciences humaines [15] les 7 et 8 novembre 2007 – dont la fonction était d'explorer l'unité et l'identité distinctive des sciences humaines – permirent de préparer le lancement effectif du programme SIPS en octobre 2011 (Figure 2).

L'idée fondatrice de la base de données était de mettre à disposition des internautes, grâce à un accès libre, des notices bibliographiques analytiques de documents en philosophie des sciences. Les spécialistes de ce domaine étaient donc invités à produire les notices des ouvrages auxquels ils ont recours dans leurs propres travaux de recherche afin de nourrir un système d'information documentaire disponible en ligne.

3.1. Dublin Core : un format de métadonnées standardisé

Le Dublin Core, développé par *Dublin Core Metadata Initiative (DCMI)*, est un formalisme de métadonnées basé sur différentes balises définissant les divers éléments d'un document. L'objectif principal de Dublin Core est de fournir un socle commun d'éléments descriptifs [10] suffisamment structuré pour permettre une interopérabilité entre des systèmes hétérogènes. Cette sémantique s'appuie sur deux ensembles distincts, *Dublin Core Element Set* et *Dublin Core Metadata Terms*. SIPS repose uniquement sur l'ensemble *Dublin Core Element Set*, qui permet une description complète des ouvrages. L'ensemble *Dublin Core Metadata Terms* est un ensemble annexe utilisé principalement pour décrire les travaux de recherche. L'ensemble utilisé dans

4. Même une base de données bibliographiques telle que *PhilPapers* (https://philpapers.org/), pourtant l'une des plus grandes au monde (sinon la plus grande actuellement) ne propose pas de notices enrichies comme en propose le SIPS.
5. https://gallica.bnf.fr/accueil/fr/
6. https://hal.archives-ouvertes.fr/
7. https://isidore.science/
8. https://logiquesagir.univ-fcomte.fr/seminaire-d-epistemologie-pratique/
9. https://www.sps-philoscience.org/les-congres/
10. https://www.dublincore.org/

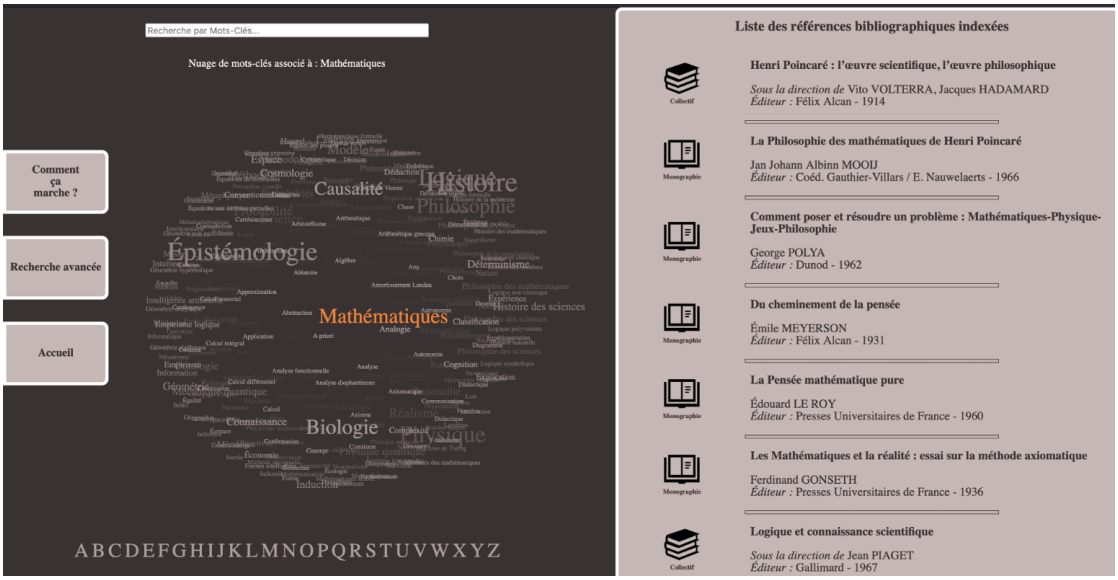


Figure 2 : Données bibliographiques du SIPS associées au mot-clé « Mathématiques ».

SIPS comprend quinze balises permettant de décrire les différents types de documents. Ces balises peuvent êtres regroupées en trois ensembles distincts : les balises relatives au contenu du document, celles relatives à sa propriété intellectuelle et enfin celles correspondant à son instanciation dans l'outil SIPS.

Tableau 1
Balises relatives au contenu.

Nom	Syntaxe	Fonctionnalité
Titre	dc:title	Indique le titre de la ressource
Créateur	dc:creator	Indique l'auteur de la ressource
Sujet	dc:subject	Indique le sujet de la ressource
Source	dc:source	Indique une référence sur une ressource dérivée (URI)
Langue	dc:language	Indique la langue de la ressource
Relation	dc:relation	Indique une référence sur une ressource apparentée
Couverture	dc:coverage	Indique le domaine d'application de la ressource

Tableau 2
Balises relatives à la propriété intellectuelle.

Nom	Syntaxe	Fonctionnalité
Créateur	dc:creator	Indique l'auteur de la ressource
Contributeur	dc:contributor	Indique les divers contributeurs de la ressource
Éditeur	dc:publisher	Indique l'éditeur de la ressource
Gestion des droits	dc:rights	Indique les droits de la ressource

Tableau 3
Balises relatives à l'instanciation.

Nom	Syntaxe	Fonctionnalité
Date	dc:date	Indique la date de création de la ressource
Type	dc:type	Indique la nature ou le genre de la ressource
Format	dc:format	Indique l'existence physique ou numérique de la ressource
Identifiant de la ressource	dc:identifier	Indique la référence unique de la ressource (URI, ISBN)

3.2. OAI-PMH : un système d'échange de collections

La plateforme SIPS propose de diffuser son contenu à l'aide de l'*Open Archives Initiative Protocol for Metadata Harvesting* [11] (OAI-PMH). Le paragraphe précédent explicitait les diverses spécifications du format de partage de métadonnées Dublin Core. Il s'agit à présent d'expliquer comment SIPS s'appuie sur le protocole OAI-PMH pour échanger son contenu. Ce protocole admet deux acteurs différents, le moissonneur en charge de récupérer des métadonnées et l'entrepôt en charge de fournir celles-ci. Il est important de noter que SIPS est un entrepôt de métadonnées destiné à être moissoné par le moteur de recherche en sciences humaines et sociales Isidore. Son objectif est de fournir au plus grand nombre les ressources philosophiques disponibles dans la base de notices bibliographiques enrichies. Le standard OAI-PMH repose sur six requêtes détaillées dans le tableau 4.

Tableau 4
Détail des requêtes OAI-PMH.

Requête	Fonctionnalité
GetRecord	Récupère un enregistrement selon un indentifiant donné.
Identify	Fournit les informations globales de l'entrepôt.
ListIdentifiers	Transmet une liste des identifiants associés aux ressources selon les paramètres donnés, à savoir, un intervalle de temps ou encore une collection de l'entrepôt.
ListMetadataFormats	Détaille au moissonneur les formats de métadonnées utilisés.
ListRecords	Transmet une liste d'enregistrements selon les paramètres donnés, à savoir, un intervalle de temps ou encore une collection de l'entrepôt.
ListSets	Indique au moissonneur, à travers une liste, les différentes collections présentes dans l'entrepôt.

4. De la liste bibliographique à la cartographie des mots-clés : la refondation du SIPS

À l'automne 2019, une nouvelle version du SIPS était mise en ligne. Dans le but de redynamiser le projet, la communauté scientifique des contributeurs était recentrée sur les membres

11. https://www.openarchives.org/pmh/

du laboratoire *Logiques de l'Agir* et les étudiants du master de philosophie de l'université de Franche-Comté. Le comité scientifique du SIPS était dès lors constitué des chercheurs du laboratoire bisontin de philosophie et de leurs partenaires au niveau national et international [12]. Il était complémenté d'un comité de pilotage composé de membres de l'établissement et créé pour assurer l'opérationnalité de la base grâce à des compétences techniques et méthodologiques mobilisées en interne [13]. Une redéfinition du périmètre de la philosophie des sciences fut alors opérée, afin d'ouvrir davantage le projet bibliographique aux sciences humaines et sociales, aux sciences médicales et à l'histoire des sciences et des techniques. Le domaine de la philosophie des sciences tel que le projet SIPS le reconcevait à partir de cette première phase de refondation pouvait dès lors recevoir un nouveau périmètre élargi autorisant une redéfinition de la philosophie des sciences comme approche réflexive de toutes les sciences.

L'objectif devenait de proposer une interface numérique à même de restituer, sous forme d'une visualisation de graphes, la richesse et la diversité des contenus déjà enregistrés de la base de données SIPS. Il fallait ainsi d'abord créer un index alphabétique de tous les mots-clés déjà enregistrés sous forme d'une barre alphabétique horizontale cliquable. Chaque lettre de la barre alphabétique devait alors renvoyer à la liste des mots-clés déjà enregistrés dans la base de données. Chaque sous-index devait renvoyer quant à lui à un graphe pondéré figurant le poids de chaque mot-clé en fonction de son taux d'indexation de notices enregistrées. Il fallait donc dans un premier temps que tout mot-clé, en tant qu'élément cliquable de l'interface graphique via la barre alphabétique, conduise au graphe d'adjacence de ses voisins dans un sous-index (Figure 3).

Figure 3 : Constellation de mots-clés du *Système d'information en philosophie des sciences* après sa refondation, 2021.

12. https://logiquesagir.univ-fcomte.fr/comite-scientifique-sips/
13. https://logiquesagir.univ-fcomte.fr/comite-de-pilotage-sips/

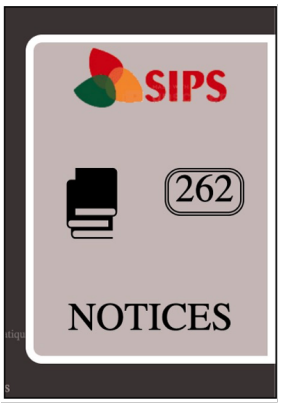


Figure 4 : Widget cliquable du SIPS indiquant le nombre de notices indexées par le mot-clé « Mathématiques ».

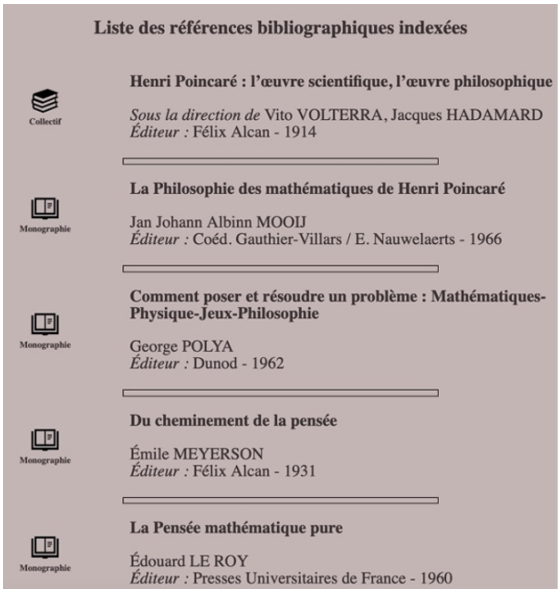


Figure 5 : Liste des références bibliographiques indexées par le mot-clé « Mathématiques » dans le SIPS.

Dans un second temps, en tant que sommet de ce graphe d'adjacence, tout mot-clé devait devenir un élément cliquable permettant d'accéder au graphe de la totalité de ses voisins dans la série des notices qu'il indexe. Chaque mot-clé, en tant que nœud, devait devenir un élément cliquable associant la liste des notices qu'il indexe dans la base de données, en proposant un widget cliquable indiquant le nombre de notices indexées (Figure 4). Ce widget, une fois cliqué, devait pouvoir proposer un élément d'interface graphique offrant la liste des documents indexés, sous forme de références bibliographiques signalétiques (Figure 5). Enfin, chaque référence signalétique devait constituer un élément cliquable de l'interface de navigation conduisant à la page de sa notice bibliographique, comprenant un abstract et ses mots-clés associés (Figure 6).

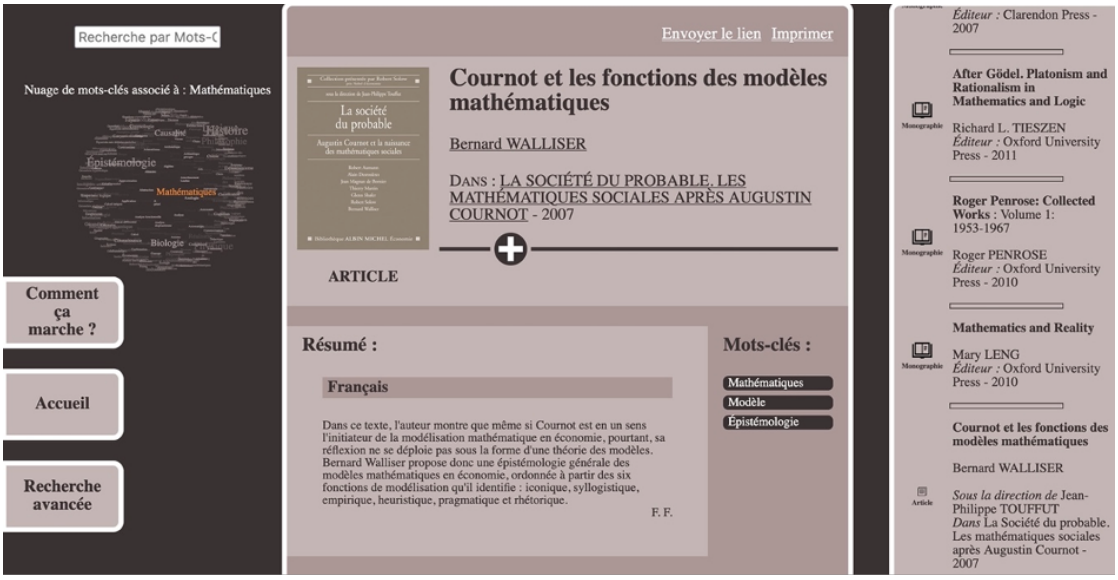


Figure 6 : Page de la notice bibliographique 2402 du SIPS indexée par le mot-clé « Mathématiques ».

En proposant un tel outil d'interface dynamique, il s'agissait ainsi de rendre possible une aide à la navigation, grâce à une interface utilisateur permettant de guider le parcours de recherche documentaire de l'internaute. Cela supplémentait cet outil de recherche analytique qu'est le moteur de recherche avancée d'une cartographie dynamique rendant la base de données SIPS interactive et plus conviviale. La finalité de cette évolution était de garantir une interface utilisateur perceptible, utilisable et compréhensible pour accroître son attractivité.

5. Du thésaurus cartographique à l'encyclopédie visuelle : l'avenir du SIPS

En s'inspirant du travail pionnier de Jack Goody [16], Pascal Robert a proposé le modèle d'une théorie générale critique des technologies intellectuelles [17] en identifiant trois grands principes (traitement de l'information, navigation documentaire et modélisation) dont il montre que l'articulation systématique, ordonnée à une opération fondamentale qu'il nomme la « conversion des dimensions », permet de penser trois grands régimes de rationalité liés à l'évolution du couple support-message : la raison graphique, la raison classificatrice et la raison simulatrice.

En régime de raison classificatrice, le traitement correspond à la mise en ordre des documents produits par l'accumulation des archives : la superposition des feuilles (documents 2D) produit en effet des volumes (documents 3D). Le traitement en raison classificatrice affronte donc le problème de la gestion, de l'organisation, de la mise en forme et du déploiement du support matériel des documents. Il conduit à une rationalisation progressive de leur organisation concrète (volumen, codex, livre), interne (mise en page, numérotation, notes, index, bibliographies, etc.) et externe (stockage dans des lieux de conservation, classification bibliothécaire).

Ce régime affronte la question de la navigation à travers le problème de l'accumulation des

volumes. Il détermine des lieux et institue des établissements (archives, bibliothèques, musées), invente des techniques de classement (classification décimale de Dewey) et des instruments de recherche (fiches, fichiers, notices, commentaires, descriptifs, catalogues, inventaires, répertoires, etc.) pour ranger les documents et y accéder. L'institution bibliothécaire est donc à la fois un territoire (c'est un lieu occupé par des bâtiments) et une carte à trois dimensions (car elle doit permettre de s'orienter dans la masse documentaire qu'elle conserve). Elle est un lieu de savoir : elle le cartographie, mais n'en révèle pas le contenu.

Dès lors c'est l'encyclopédie qui modélise en régime de raison classificatrice, car son projet infini est d'appréhender le savoir dans son épaisseur et dans son mouvement, dans sa dynamique spatio-temporelle et spatio-culturelle. Elle vise à rassembler les savoirs dans une unité organique et à restituer dans leur essence et leur concision les connaissances humaines.

Mais lorsque l'encyclopédie entre dans le régime de la « raison simulatrice », lorsqu'elle se numérise et s'interface, elle acquiert des caractères inédits qu'elle n'avait encore jamais eu auparavant, comme en témoigne l'encyclopédie collaborative *Wikipedia* [18]. Elle devient mondiale et foncièrement démocratique, elle rend possible un accès direct aux sources des notices, à leur révision et enrichissement perpétuel, à leur lecture comparée et multilingue. Enfin elle offre une dynamicité des liens (hypertextualité) et des contenus (simulation numérique), ainsi qu'une interactivité (qui est un corollaire de l'hypertextualité et de la simulation).

Dès lors si la connaissance signifie la capacité à effectuer une action pour atteindre un but fixé [7, p. 77], alors une conception imagée des connaissances consiste à les figurer à travers des dessins pourvus de sens. Ce sens doit être la transparence de ce que ces dessins visent objectivement, à la manière d'une photographie vis-à-vis de ce dont elle est la photographie : à savoir son objet.

Reposer la question encyclopédique à partir d'une conception imagée de la connaissance à l'ère des interfaces dynamiques, c'est donc poser la question de l'enchaînement des connaissances à travers leurs « dessins » ou schématisations. C'est donc aussi poser la question de l'articulation de ces dessins comme suites de gestes à enchaîner afin de réaliser une tâche plus ou moins complexe. La figuration de cet enchaînement introduit la question de la grammaire permettant d'articuler ces schémas les uns aux autres pour dire le sens de la tâche plus ou moins complexe qu'ils signifient et permettent de viser.

Dès lors on peut distinguer dans le projet de constitution d'une encyclopédie visuelle deux dimensions du langage graphique de figuration qu'elle doit mobiliser : celle du répertoire de symboles élémentaires constituant son vocabulaire visuel et celle des règles de composition de ces éléments constituant sa grammaire visuelle [19]. En déterminant une telle boîte à outils symboliques et le mode d'emploi qui lui est associé, il devient possible de construire une encyclopédie visuelle de la connaissance pratique. Telle pourrait devenir la mission future du SIPS parallèlement à la cartographie numérique des sciences qu'il propose actuellement à travers des constellations de mots-clés qui permettent d'accéder à ses notices bibliographiques enrichies.

Conclusion : pratiques scriptovisuelles et images opérationnelles

La nature des infrastructures réseaux actuelles, les possibilités ouvertes par l'internet des objets et la réalité augmentée induisent une mutation des formes de la lecture et de l'écriture sur écrans. Elles permettent des formes d'interaction qui dépassent le simple accès à des contenus via des liens hypertextuels.

Depuis Harun Farocki, on parle en effet d'« images opérationnelles » (*operational images*) [20]. Documentariste et vidéaste, Farocki s'est concentré sur la politique de l'imagerie dans le complexe militaro-industriel pour montrer que la fonction de cette classe d'images n'est ni de représenter ni même d'informer, mais de surveiller, de détecter et d'identifier à distance [14] grâce aux nouveaux dispositifs de télé-action. Elles sont constitutives de ce que Grégoire Chamayou a récemment appelé des « sociétés de ciblage » [21].

Les images opérationnelles sont des interfaces qui permettent d'exécuter des tâches dans le cadre de processus opérationnels. Elles comprennent diverses technologies d'imagerie qui associent souvent des caméras ou des capteurs à des logiciels de traitement d'images (véhicules aériens sans pilote, voitures autonomes, robots industriels et domestiques, imagerie médicale, scanners industriels, systèmes d'information géographique, etc). Elles permettent de poser des questions sur le statut des images à l'époque des écrans et des technologies visuelles. Dès lors, en interfaçant cartographie des sciences et encyclopédie visuelle d'une part, pratiques scriptovisuelles et images opérationnelles d'autre part, le SIPS pourrait devenir une interface homme-machine et un système d'apprentissage nouveau déployé en réalité virtuelle augmentée.

Références

[1] G. Frege, Les fondements de l'arithmétique : recherche logico-mathématique sur le concept de nombre, L'ordre philosophique, Seuil, 1969.

[2] S. Schreibman, R. G. Siemens, J. Unsworth (Eds.), A Companion to Digital Humanities, number 26 in Blackwell companions to literature and culture, Blackwell Publishing Ltd, 2004.

[3] B. Bachimont, Du texte à l'hypotexte : les parcours de la mémoire documentaire, in : Mémoire de la technique et techniques de la mémoire, Technologies, idéologies, pratiques, Érès, 1999, pp. 195–225. URL : https://www.cairn.info/memoire-de-la-technique-et-techniques--9782865866557-p-195.htm. doi :10.3917/eres.lenay.1999.01.0195.

[4] B. Stiegler, Machines à écrire et matières à penser, Genesis (Manuscrits-Recherche-Invention) 5 (1994) 25–49. URL : https://www.persee.fr/doc/item_1167-5101_1994_num_5_1_952. doi :10.3406/item.1994.952, publisher : Persée - Portail des revues scientifiques en SHS.

[5] B. Stiegler, Le carnaval de la nouvelle toile : de l'hégémonie à l'isonomie, in : B. Juanals, J.-M. Noyer (Eds.), Technologies de l'information et intelligences collectives, Systèmes d'information et organisations documentaires, Hermès, 2010.

14. https://operationalimages.cz/

[6] B. Stiegler, Technologies de la mémoire et de l'imagination, Réseaux. Communication - Technologie - Société 4 (1986) 61–87. URL : https://www.persee.fr/doc/reso_0751-7971_1986_num_4_16_1204. doi :10.3406/reso.1986.1204, publisher : Persée - Portail des revues scientifiques en SHS.

[7] B. Bachimont, Arts et sciences du numérique : ingénierie des connaissances et critique de la raison computationnelle, Université de technologie de Compiègne, 2004. Mémoire d'HDR.

[8] B. Bachimont, Bibliothèques numériques audiovisuelles : des enjeux scientifiques et techniques, Document Numérique 2 (1998) 219–242.

[9] G. Berry, Pourquoi et comment le monde devient numérique : leçon inaugurale prononcée le jeudi 17 janvier 2008, number 197 in Leçons inaugurales du Collège de France, Collège de France / Fayard, 2008.

[10] G. Berry, Penser, modéliser et maîtriser le calcul informatique : leçon inaugurale prononcée le jeudi 19 novembre 2009, number 208 in Leçons inaugurales du Collège de France, Collège de France / Fayard, 2009.

[11] B. Bachimont, Le contrôle dans les systèmes à base de connaissances : contribution à l'épistémologie de l'intelligence artificielle, Hermès, 1994. 2ème édition.

[12] A. M. Turing, On Computable Numbers, with an Application to the Entscheidungsproblem, Proceedings of the London Mathematical Society s2-42 (1937) 230–265. URL : http://doi.wiley.com/10.1112/plms/s2-42.1.230. doi :10.1112/plms/s2-42.1.230.

[13] J. Tomm, G. Leroux, La collection Raymond Klibansky conservée à l'Université McGill : présentation de la bibliothèque d'un humaniste montréalais, Mémoires du livre 5 (2013). URL : http://id.erudit.org/iderudit/1020226ar. doi :10.7202/1020226ar.

[14] J.-L. Pinol, Une infrastructure pour les SHS : le TGE adonis, Revue d'histoire moderne et contemporaine 58-4bis (2011). URL : http://www.cairn.info/revue-d-histoire-moderne-et-contemporaine-2011-5-page-90.htm. doi :10.3917/rhmc.585.0090.

[15] T. Martin (Ed.), Les sciences humaines sont-elles des sciences ?, Philosophie des sciences, Vuibert, 2011.

[16] J. Goody, La raison graphique : la domestication de la pensée sauvage, Le sens commun, Les Éditions de Minuit, 1979.

[17] P. Robert, Mnémotechnologies : une théorie générale critique des technologies intellectuelles, Communication, médiation et construits sociaux, Hermès science publications : Lavoisier, 2010.

[18] L. Barbe, M. Severo (Eds.), Wikipédia, objet de médiation et de transmission des savoirs, Intelligences numériques, Presses universitaires de Paris Nanterre, 2021.

[19] P. Descola, Les formes du visible : une anthropologie de la figuration, Les livres du nouveau monde, Éditions du Seuil, 2021.

[20] V. Pantenburg, Working images : Harun Farocki and the operational image, in : J. Eder, C. Klonk (Eds.), Image Operations : Visual media and political conflict, Manchester University Press, 2016.

[21] G. Chamayou, Avant-propos sur les sociétés de ciblage : une brève histoire des corps schématiques, Jef Klak (2015-09-21). URL : https://www.jefklak.org/avant-propos-sur-les-societes-de-ciblage/.

Représenter et étudier les individus dans un corpus numérique : le cas de la correspondance d'Henri Poincaré

Representing and studying people in a digital corpus: the case of the Henri Poincaré correspondence

Nicolas Lasolle[1,2,*], Laurent Rollet[1]

[1]*Université de Lorraine, CNRS, Université de Strasbourg, AHP-PReST, F-54000 Nancy, France*
[2]*Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France*

Abstract
Henri Poincaré (1854-1912) is a French scientist known for his major contributions to the fields of mathematics, physics and philosophy of science. His correspondence consists of more than 2000 letters, exchanges involving numerous correspondents from his private and professional sphere. Within the *Archives Henri-Poincaré* laboratory, several works are dedicated to the study and the valorization of this historical corpus ; with a particular interest in digital publishing and in the proposal of advanced search tools relying on Semantic Web technologies. This article focuses on the question of individuals within a historical corpus, based on the case of this correspondence corpus. Methodological aspects relating to the numerical representation of actors and their relationships are presented. We then discuss methods and tools dedicated to the exploration of the corpus. In particular, recent work has led to the creation of a flexible querying system which can facilitate the discovery of connections between items in the corpus and which can support the identification of networks of people. By presenting historical researches focused on actors, we try to show the interest of digital approaches, in particular to support a study of the corpus from below, which brings to the fore actors and institutions usually considered as secondary.

Keywords
history, digital humanities, Semantic Web, knowledge base, flexible querying

1. Introduction

Henri Poincaré est né à Nancy en 1854, et est décédé à Paris en 1912. Tour à tour étudiant de l'École polytechnique (1873-1875) et de l'École des mines (1875-1878), il a obtenu un doctorat ès sciences mathématiques en 1879 à la faculté des sciences de Paris et s'est rapidement inscrit au sein d'un réseau scientifique et administratif important, comme en témoigne sa correspondance active et passive. Constitué d'environ 2200 lettres et allant de son entrée à l'École polytechnique en 1873 jusqu'à son décès en 1912, ce corpus témoigne de l'évolution professionnelle et personnelle de ce scientifique. L'étude et la valorisation de ce corpus est un projet de longue date qui

. *Workshop on Digital Humanities and Semantic Web*
*. Corresponding author.
. nicolas.lasolle@univ-lorraine.fr (N. Lasolle) ; laurent.rollet@univ-lorraine.fr (L. Rollet)
. https://github.com/nlasolle (N. Lasolle) ;
https://poincare.univ-lorraine.fr/fr/membre-titulaire/laurent-rollet (L. Rollet)
. 0000-0002-1253-649X (N. Lasolle)

a conduit à la publication de quatre volumes thématiques. Le premier regroupe les échanges avec le mathématicien suédois Gösta Mittag-Leffler [1], échanges réguliers s'étalant durant la majeure partie de la carrière académique d'Henri Poincaré [1]. Le deuxième volume concerne les échanges avec des physiciens, des chimistes et des ingénieurs, [2]. Le troisième constitue la correspondance entre Henri Poincaré, des astronomes, et les géodésiens [3]. Le volume le plus récent publié à ce jour est dédié à la correspondance de jeunesse d'Henri Poincaré [4] lorsqu'il était en formation à l'École polytechnique puis à l'École des mines (1873-1878). Deux autres volumes sont en préparation au moment de la rédaction de ce document et devraient clôturer ce projet d'édition. L'un est dédié à la correspondance avec les mathématiciens et l'autre concerne la correspondance administrative et privée.

L'étude de ce corpus historique présente des enjeux scientifiques variés, de par les nombreux thèmes abordés au sein des échanges tenus par Poincaré [5]. Tout d'abord, l'étude d'une correspondance telle que celle d'Henri Poincaré permet d'apporter des éléments de contexte relatifs à l'élaboration de théories scientifiques. En effet, à la fin du XIX^{e} siècle, la lettre était un vecteur non négligeable de l'information scientifique [6, 7]. Au cours de sa carrière, Henri Poincaré a tenu une correspondance active avec nombre de ses confrères français et étrangers. Parmi ces correspondants scientifiques, nous pouvons citer le mathématicien suédois Gösta Mittag-Leffler, l'Italien Giovanni Battista Guccia ou encore le mathématicien allemand Felix Klein. L'étude de la correspondance d'Henri Poincaré peut également éclairer le fonctionnement de différentes institutions et sociétés savantes de son époque de par les échanges réguliers qu'il tenait avec plusieurs autres membres. Explorer cette correspondance permet également de mieux comprendre le contexte politique, social et culturel de la France de la fin du XIX^{e} siècle. Enfin, ce corpus présente Henri Poincaré sous un jour différent de ce qui est mis en avant par l'étude de ses travaux et permet ainsi d'introduire de nouveaux éléments biographiques. Ces échanges épistolaires peuvent révéler plusieurs points de la personnalité de Poincaré relatifs à sa méthode de recherche, ses relations, sa vie privée, etc.

Le corpus rassemble de nombreux acteurs, appartenant au cercle privé et professionnel de Poincaré. On dénote environ 500 correspondants et plus de 1000 personnes citées. Ces individus sont majoritairement français, mais de nombreux acteurs de premier plan sont des scientifiques ou académiciens étrangers tels que Gösta Mittag-Leffler, Felix Klein, ou Giovanni Battista Guccia. Dans le cadre de travaux historiques, de nombreuses recherches peuvent également amener les chercheurs à s'intéresser à des acteurs généralement considérés comme secondaires. Nous proposons de discuter des éléments relatifs au concept d'histoire par en bas avant de présenter le contexte numérique des travaux relatifs au corpus de la correspondance d'Henri Poincaré. En particulier, nous présenterons comment des outils numériques peuvent aider les chercheurs menant des recherches sur le corpus en s'appuyant sur les acteurs. Dans ce contexte, nous présenterons des travaux relatifs à la formalisation et à l'utilisation d'un système d'interrogation flexible pour le Web sémantique. Sans détailler les aspects techniques, nous proposons d'illustrer son fonctionnement et son intérêt au travers de plusieurs exemples issus du corpus de la correspondance d'Henri Poincaré. Ce système a notamment été intégré à un outil de navigation

1. Gösta Mittag-Leffler (1846-1927) est notamment reconnu pour avoir introduit un théorème (portant désormais son nom) relatif à la représentation des fonctions méromorphes par des séries de fractions rationnelles. En 1882, il crée la revue *Acta Mathematica* où publieront plusieurs mathématiciens de l'époque comme Georg Cantor, Sofia Kovalevskaïa et Henri Poincaré.

permettant une exploration interactive du corpus.

2. Du mode d'existence des invididus dans le corpus poincaréien

Le projet d'édition de la correspondance de Poincaré a été fondé dans les années 1990 en faisant le choix de publier l'ensemble des lettres, c'est-à-dire à la fois les lettres relevant de l'activité scientifique et intellectuelle du mathématicien et celles relevant de la sphère privée et administrative. Une telle décision n'allait pas forcément de soi à cette époque. Ce corpus de plus de 2000 lettres contenait en effet — outre des échanges épistolaires avec des grands noms de la science et de la scène intellectuelle française et étrangère — plusieurs centaines de lettres d'acteurs peu connus ou peu étudiés voire totalement inconnus : des membres de la famille élargie, des amis d'enfance, des amateurs, des journalistes, des inventeurs, des fous littéraires ou scientifiques, etc. Parallèlement à ce travail éditorial les chercheurs ont exploré systématiquement tous les services d'archives susceptibles de contenir des lettres envoyées par Poincaré afin de reconstituer au mieux des échanges épistolaires homogènes [2]. Ces recherches ont ainsi permis de retrouver aux Archives nationales près d'une centaine de traces d'échanges entre Poincaré et les différents ministères dont il a dépendu durant sa carrière.

Que faire de ces documents parfois très riches en termes de contenus mais difficiles à exploiter en raison même des difficultés d'identification des acteurs et de leur mise en situation dans un contexte temporel ou thématique ? En effet de nombreuses lettres reçues par Poincaré n'étaient ni signées ni datées, certaines avaient même été retrouvées sous la forme de projets de brouillons dans les dossiers de carrière de Poincaré sans que l'on puisse disposer de la lettre définitive reçue par lui. De plus, lorsque Poincaré s'adressait, par exemple, au ministre des Travaux publics il était facile de déterminer l'identité du ministre destinataire mais, dans la mesure où la lettre de réponse était rédigée par un fonctionnaire de son cabinet et qu'on ne connaît pas le nom de son auteur, se posait alors la question de savoir s'il fallait attribuer cette lettre au ministre lui-même. La tentation était grande de ne faire porter les efforts de recherche et d'édition que sur les acteurs « intéressants » ayant déjà fait l'objet d'une forme de patrimonialisation historique (les savants de premier ordre, les ministres, les membres de l'Académie française). Le choix a finalement été tout autre : décision a été prise de prendre en compte toutes les lettres quels que soient leurs auteurs ou leurs contenus pour restituer l'image la plus complète possible des travaux, de la personnalité et du rayonnement social, public et mondain du mathématicien. Une telle démarche revenait non pas à considérer que tous les correspondants devaient être mis sur un même pied d'égalité mais qu'ils devaient au moins bénéficier d'une égale considération en termes d'analyse et de traitement éditorial. Plutôt que de se concentrer sur quelques acteurs majeurs pour rendre compte de la vie et de l'œuvre de Poincaré — Paul Appell, Pierre et Marie Curie, Lazarus Fuchs, Gösta Mittag-Leffler, Paul Painlevé pour n'en citer que quelques-uns — nous avons préféré une approche ouverte dans laquelle Poincaré était considéré comme le centre d'un réseau épistolaire dans lequel il pouvait compter d'une manière ou d'une autre pour ses correspondants et dans lequel ces acteurs pouvaient également avoir de l'importance pour

2. Le corpus contient en effet plus de 1000 lettres écrites de la main de Poincaré et près de 1000 qui lui été adressées.

lui. Cette démarche prenait appui sur un parti pris méthodologique en phase avec l'évolution récente de l'historiographie en histoire des sciences.

Avant de se constituer en tant que discipline au début du xx^{e} siècle, l'histoire des sciences s'est d'abord construite sur le mode des éloges académiques et funéraires des grands savants. Et pendant longtemps elle a gardé de cette filiation une propension très forte à ancrer ses sujets d'études sur les acteurs de premier plan et à se focaliser essentiellement sur leurs activités relevant des sciences. L'ancrage social, voire politique des acteurs, leur inscription dans des réseaux scientifiques, intellectuels et mondains, la diversité des processus d'institutionnalisation des sciences ont souvent été écartés au profit d'approches *internalistes* basées, implicitement ou explicitement, sur le postulat que le travail scientifique peut être analysé indépendamment (ou presque) des contextes sociaux, économiques ou politiques qui le rendent possible. L'émergence des approches sociales et sociologiques des sciences et techniques a partiellement rebattu les cartes méthodologiques. Depuis quelques décennies, les historiens des sciences ont profondément revisité leurs méthodes et leurs sujets d'études : afin de mieux articuler approches internalistes et externalistes ils sont devenus très sensibles à la l'élaboration d'une *histoire par en bas* ou d'une *histoire populaire* des sciences qui donne la une place plus équilibrée aux acteurs secondaires : artisans, techniciens, ingénieurs, amateurs, acteurs professionnels des sciences, enseignants, militaires des armes savantes, enseignants, étudiants, etc. [3]

L'intérêt de cette approche est qu'elle invite à faire un pas de côté et à dépasser l'image parfois caricaturale de la *Science* au profit de la perspective plus large des *savoirs scientifiques en sociétés* ; elle invite également à décentrer le regard en considérant avec sérieux des acteurs qui bien qu'ils n'aient pas toujours bénéficié d'un fort capital académique ou scientifique ont pu jouer un rôle non négligeable dans le champ scientifique et intellectuel [4].

La mise en œuvre de ce type d'approche dans le cadre de l'édition de la correspondance de Poincaré n'est pas sans conséquences sur le travail à réaliser. Tel qu'il a été conçu, que ce soit dans les ouvrages publiés ou en cours d'édition ou sur la plate-forme numérique dédiée, ce projet vise à « épuiser » les contextes de Poincaré et d'élaborer une encyclopédie totale sur cet acteur de premier plan en exploitant toutes les traces et sources disponibles [5]. Il est conçu dans une logique de valorisation et de vulgarisation scientifique avec pour perspective finale l'élaboration d'une biographie du mathématicien. Pour ce faire, un important travail doit être réalisé sur chacun des individus — correspondants, personnes citées dans les lettres et dans les apparats critiques — pour caractériser *a minima* leur identité sociale et professionnelle et déterminer les modalités de leurs relations avec Poincaré. La base de connaissances contient en effet des milliers de noms d'acteurs qui peuvent être des correspondants (pas toujours facilement identifiables), des amis, des « ennemis », des relations familiales, de vagues connaissances, des inconnus (non

3. Un bon exemple de cette approche est le livre de Clifford D. Conner *Histoire populaire des sciences* [8]. Celui-ci s'inspire d'ailleurs des approches développées en histoire générale, notamment par Howard Zinn dans son *Histoire populaire des Etats-Unis. De 1492 à nos jours* [9]. Pour une approche similaire en France, voir le livre de Gérard Noiriel *Histoire populaire de la France, de la Guerre de Cent Ans à nos jours* [10].

4. On ne donnera ici que quelques exemples représentatifs. La Société astronomique de France créée en 1882 par l'astronome et vulgarisateur Camille Flammarion a joué un rôle essentiel pour la structuration des sciences astronomiques mais elle comptait parmi ses membres, outre les grands noms de ces disciplines, des centaines d'amateurs ou de non professionnels, dont un nombre significatif de femmes.

5. Le site Web de la correspondance n'est d'ailleurs qu'une partie d'un écosystème numérique qui inclut d'autres plates-formes dédiées à ses publications, à sa biographie et aux documents iconographiques le concernant.

seulement des historiens mais de Poincaré lui-même), des collègues, des personnalités politiques, des philosophes, des acteurs de l'histoire politique ancienne ou récente, des auteurs d'œuvres littéraires [6].

Pour ne donner qu'un exemple, la lettre relativement anodine de Poincaré ci-dessous, datant probablement de mars 1874, fait le récit d'une conversation lors d'un voyage en train de Nancy vers Paris avec des camarades de promotion à l'École polytechnique. Elle porte sur la beauté de jeunes femmes de la bonne société nancéienne puis sur les classements intermédiaires après les examens de février 1874. On y trouve près d'une vingtaine de références à des individus. Pour chacun d'entre eux, il a été nécessaire d'effectuer des recherches approfondies afin de les identifier et de déterminer la nature du lien relationnel qu'ils entretenaient avec Poincaré. On constate la présence d'au moins 5 strates de personnes qui renvoient à différents univers sociaux : des camarades polytechniciens ayant fait leurs études avec Poincaré au lycée de Nancy (les 2 premières lignes de la lettre), un groupe de 4 jeunes femmes nancéiennes, parfois aristocrates, qui renvoient à l'univers amical [7], familial et mondain de la famille Poincaré à Nancy (lignes 2 et 3), une référence à un très proche ami d'enfance alors élève à l'École militaire de Saint-Cyr (Élie Rinck, lignes 3 et 4), un ensemble de camarades de promotion de l'École polytechnique (second paragraphe de la lettre) et une référence à une dame nancéienne, Mme Dreyfuss, qui conduisait le cotillon (le bal), lors d'une soirée mondaine. En outre, au-delà des noms évoqués par Poincaré, cette lettre fournit d'intéressants indices sur la personnalité et la psychologie de Poincaré dans sa jeunesse et met bien en évidence l'univers social et mondain dans lequel il évoluait alors.

La question qui se pose à l'historien éditeur de la lettre est alors de déterminer de quelle manière caractériser toutes ces personnes, pour autant qu'il soit possible de le faire. Cela revient à se poser des questions que seules une approche biographique et une connaissance approfondie des sources biographiques peuvent éclairer : ces personnes comptaient-elles pour Poincaré et à quel titre ? Quel était le degré de proximité de Poincaré avec elles ? Poincaré comptait-il réciproquement pour ces personnes ? S'agissait-il de personnes sans importance dans son vécu social ou jouaient-elles un rôle quelconque dans son parcours de vie ? Se pose également la question des sources mobilisables : il est possible de s'appuyer sur des données « objectives », à savoir d'autres lettres, des réseaux de citations, des photographies, des documents d'archives, des récits familiaux comme le journal rédigé par la sœur de Poincaré [11]. Il est également possible d'exploiter des données « indirectes », sujettes à des interprétations et à des hypothèses, telles que les connexions généalogiques identifiables, d'autres fonds de correspondance, des témoignages de seconde main, des bases de données en ligne telles que celle recensant tous les élèves passés par l'École polytechnique. Mais la grande difficulté est d'équilibrer l'investissement en termes de recherche d'informations en fonction de l'importance réelle ou supposée de ces personnes vis-à-vis de Poincaré. Dans le cadre de cette lettre, il apparaît qu'au regard de la biographie poincaréienne, les noms de Rinck, Bonnefoy, Petitdidier et Corps sortent nettement du lot car les sources épistolaires attestent l'existence d'une relation suivie entre Poincaré et ces

6. Ainsi durant ses années d'études à l'École polytechnique et à l'École des mines (1873-1878), Poincaré allait très souvent au théâtre et à l'opéra et il adorait les pièces d'opéra-bouffe ou les vaudevilles. Sa correspondance de jeunesse abonde de références aux auteurs de ces œuvres ainsi qu'aux comédiens et comédiennes qu'il voyait régulièrement sur scène [4]

7. Mademoiselle Jacquinet était ainsi la fille du recteur de l'Académie de Nancy et une amie proche de la sœur de Poincaré [11].

Figure 1 : Transcription d'une lettre d'Henri Poincaré à Eugénie Poincaré, telle que présentée sur le site http://henripoincare.fr.

personnes.

Face à ces milliers de personnes l'enjeu est donc de constituer une base de connaissances qui permette d'identifier ces individus et de les caractériser en termes sociaux sans tomber dans le piège de l'exhaustivité. Il ne s'agit donc pas d'élaborer des biographies individuelles précises — les biographies sur les grands correspondants de Poincaré ne manquent pas — mais de leur attribuer *a minima* une identité personnelle, sociale et professionnelle en n'oubliant pas que cette caractérisation a pour destination l'explicitation du corpus poincaréien. Cela revient ainsi à définir ces acteurs par rapport à une typologie de métiers (professeur, homme politique, actuaire, ingénieur, etc.), à des attaches institutionnelles (École polytechnique, Académie des sciences, Académie des sciences, Banque de France, Société mathématique de France, etc.) ou à des appartenances disciplinaires (mathématicien, physicien, philosophe, etc.). Une telle entreprise peut sembler relativement simple mais peut s'avérer complexe.

Comment ainsi caractériser l'épouse de Poincaré, Louise Poulain d'Andecy (1857-1834) ? La facilité serait sans doute de la définir comme « femme au foyer et épouse de Poincaré » car elle n'eut manifestement aucune activité professionnelle [12]. Mais les sources nous indiquent qu'elle fut tout à la fois pour Poincaré sa secrétaire, sa « chargée de relations » et la

relectrice et correctrice de ses manuscrits (tout comme d'ailleurs ses enfants). Dans le même ordre d'idée, caractériser les attaches professionnelles de Poincaré, ou d'un très grand nombre de ses correspondants, s'avère complexe dans la mesure où beaucoup étaient membres de plusieurs dizaines d'académies et de sociétés savantes [8]. Par ailleurs, s'agissant des disciplines, un problème similaire se pose : faut-il caractériser Poincaré comme mathématicien, comme philosophe, comme physicien, comme astronome, comme administrateur de la recherche ? Tous ses qualificatifs peuvent s'appliquer à son parcours de vie et de carrière et ils posent la question redoutable de la profondeur temporelle que l'on veut attribuer à la base de connaissances des individus — Poincaré n'était pas mathématicien à 2 ans et il n'a publié des articles relevant de la philosophie qu'à partir du début des années 1890. Enfin, le statut des femmes est à considérer avec minutie : Eugénie Launois (1830-1897) était la mère de Poincaré et de sa sœur Aline ; elle était également l'épouse du père de Poincaré, Émile Léon Poincaré (1828-1890) et avait donc pour nom de mariage « Eugénie Poincaré ». Cette simple caractérisation ouvre une infinité de possibles : Eugénie Launois *alias* Eugénie Poincaré *alias* l'épouse du père de Poincaré *alias* la femme d'un professeur à la faculté de médecine de Nancy *alias* la mère de Poincaré *alias* la belle maman de Louise Poulain d'Andecy (l'épouse de Poincaré) *alias* la maman d'Aline Poincaré, sœur de Poincaré et future épouse du philosophe Émile Boutroux, une influence philosophique majeure pour Poincaré…

Pour finir, le traitement des individus dans un corpus de correspondance s'avère complexe lorsqu'il s'agit de les intégrer dans une base de connaissances. Il oblige à mener en amont une réflexion sur la notion d'identité sociale et sur ses transformations à travers le temps de vie d'un acteur. Il nécessite aussi d'analyser différents modes d'existence des acteurs : ce qu'ils sont par et pour eux-mêmes, ce qu'ils sont par et pour les autres, ce qu'ils déclarent / croient / aimeraient être, ce qu'ils sont au sein (ou en dehors) d'un groupe social, ce que les chercheurs et historiens déclarent ce qu'ils sont, ce qu'ils sont dans une représentation centrée sur le corpus poincaréien, ce qu'il sont lorsqu'on les caractérise à une époque spécifique en lien avec un échange épistolaire spécifique. Et ce ne sont que quelques possibles dont l'historien et l'informaticien doivent tenir compte. Dans la suite de cet article, nous proposons de détailler les approches numériques mises en place pour l'exploitation de ce corpus.

3. L'édition numérique

Depuis plusieurs années, les Archives Henri-Poincaré mènent différents travaux numériques autour du corpus de la correspondance d'Henri Poincaré. Le site Web utilisé pour éditer et publier le corpus de la correspondance d'Henri Poincaré est géré grâce au gestionnaire de contenus (CMS) Omeka S [9] [13]. Ce système permet la mise en valeur numérique de collections liées au patrimoine culturel, qu'elles soient issues de musées, bibliothèques ou centres d'archives.

Le site Web lié au corpus est accessible à l'adresse http://henripoincare.fr, et se compose de 4 sous-sites : l'un dédié au corpus de la correspondance ; un deuxième qui s'articule autour de la bibliographie sur et par Poincaré ; un troisième qui présente des éléments biographiques relatant le parcours de Poincaré et un dernier proposant une iconographie regroupant des dizaines

8. Au cours de sa carrière Poincaré fut élu dans 45 académies françaises et étrangères.
9. https://omeka.org/s/.

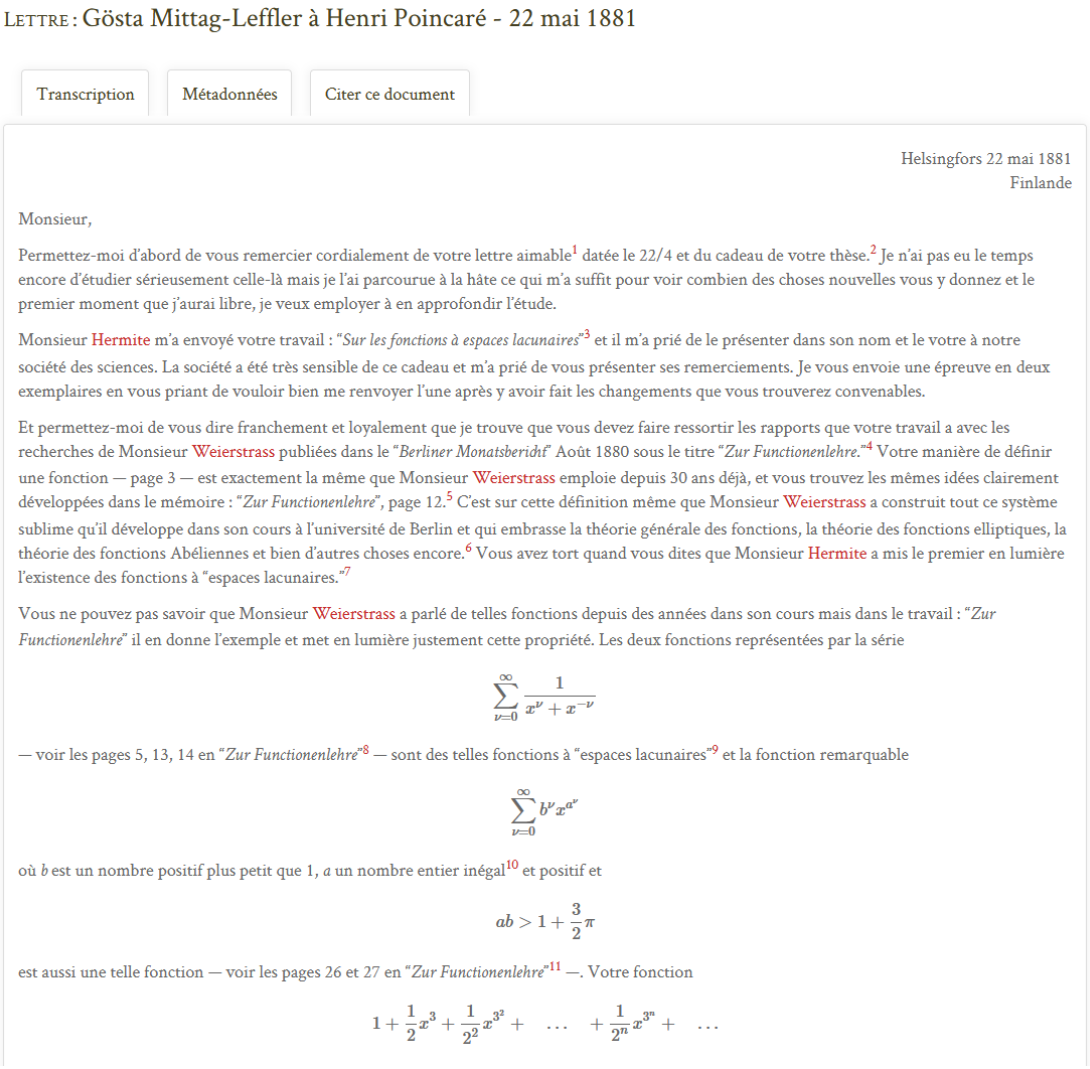


Figure 2 : Extrait de la transcription d'une lettre telle que présentée sur le site henripoincare.fr.

d'images présentant Poincaré, sa famille, ses collègues, etc. Cet ensemble de sites Web est hébergé par Huma-Num [14] qui propose des services adaptés aux projets d'humanités numériques. La figure 2 présente une capture du site, correspondant à un extrait de la transcription d'une des premières lettres envoyées par le mathématicien suédois Gösta Mittag-Leffler à Henri Poincaré, le 22 mai 1881. Cette lettre traite notamment des travaux de thèse effectuées par Poincaré et des liens soulignés par Mittag-Leffler avec les travaux du mathématicien allemand Karl Weierstrass, l'une des figures majeures de « l'École de Berlin », qui a apporté des contributions fondamentales dans plusieurs domaines tels que la théorie des fonctions analytiques et le calcul des variations.

Omeka S permet de représenter les données sous la forme de triplets dans la logique du

```
⟨lettreA expéditeur henriPoincaré⟩
⟨lettreA destinataire marieSklodowskaCurie⟩
⟨lettreA rédigéÀ Paris⟩
⟨lettreA thème "Prix nobel"⟩
⟨lettreA dateDeRédaction 1911-12-11⟩
```

Figure 3 : Un exemple de triplets décrivant partiellement une lettre.

modèle RDF [10] [15], modèle principal pour représenter les données du Web sémantique. Un ensemble de triplets RDF forme un graphe composé de trois types de nœuds : des *ressources nommées*, des *ressources anonymes* et des *littéraux*. Une *ressource nommée* est identifiée par un IRI et permet de décrire une classe (p. ex. `Personne`, `Mathématicien`, `Lettre`, etc.), une propriété (p. ex. `expéditeur`, `destinataire`, etc.) ou une instance (p. ex. `henriPoincaré`, `lettre11`, etc.) [11]. Une *ressource anonyme* représente une ressource qui n'est pas explicitement identifiée (*nœud vide*). Un littéral correspond à une valeur constante d'un type donné (entier, chaîne de caractères, date, etc.), c'est un objet atomique et primitif.

Il est possible de définir des relations entre les nœuds du graphe par l'utilisation de propriétés décrivant les ressources les composant. Ces relations sont caractérisées par des triplets de la forme ⟨*sujet prédicat objet*⟩. Le sujet représente la ressource (nommée ou anonyme) à décrire. Le prédicat est une propriété qui décrit cette ressource. L'objet est la valeur associée à la propriété et peut être une ressource nommée, une ressource anonyme ou un littéral. La figure 3 donne un exemple de description partielle d'une lettre de la correspondance d'Henri Poincaré.

Bien qu'Omeka soit un outil s'inscrivant dans le mouvement du Web sémantique [12] il présente plusieurs limites [16]. En effet, il n'est pas possible d'intégrer plusieurs éléments introduits par le langage RDFS tels que des hiérarchies entre des classes (par ex. une lettre autographe est une forme de lettre qui est une forme de document) et entre des propriétés (un destinataire est un correspondant). Ce type de relations est le support de mécanismes d'inférences permettant de tirer parti des connaissances associées à un domaine. De plus, dans le contexte du Web sémantique, le langage SPARQL permet de formuler des requêtes expressives afin d'exploiter les liens entre les ressources [17]. Les données Omeka S sont stockées à l'aide d'une base de données MySQL dédiée qui ne peut pas être directement interrogée par des requêtes SPARQL.

Ces différents éléments ont conduit à la création d'une base au format RDF, alimentée quotidiennement par un export et une conversion des données du site Omeka S [13]. Les données de cette base sont exposées au travers d'un point d'accès SPARQL, qui permet l'interrogation du graphe RDF. Par exemple, la requête `Q` (représentée de façon informelle avec $\mathcal{Q}$) peut être formulée pour interroger cette base.

10. *Resource Description Framework.*

11. Dans un souci de lisibilité, les ressources nommées ne sont pas représentées en utilisant des IRI complets dans cet article.

12. Omeka S est une version « sémantique » du logiciel Omeka, qui propose l'intégration d'ontologies et la possibilité de lier des ressources, comme le fait le modèle RDF.

13. Ce script, développé avec le langage Python, est générique et peut être réutilisé pour toute base Omeka S. Le code source et des détails sur son utilisation sont disponibles sur un dépôt GitHub public (https://github.com/nlasolle/omekas2rdf).

$$Q = \left| \begin{array}{l} \texttt{SELECT ?}\ell \\ \texttt{WHERE \{} \\ \quad \texttt{?}\ell\ \texttt{a Letter .} \\ \quad \texttt{?}\ell\ \texttt{sentBy henriPoincaré .} \\ \quad \texttt{?}\ell\ \texttt{cite ?x .} \\ \quad \texttt{?x a Mathematician .} \\ \quad \texttt{?}\ell\ \texttt{subject "geometry"} \\ \texttt{\}} \end{array} \right.$$

$$\mathcal{Q} = \left| \begin{array}{l} \text{Donner les lettres envoyées par Henri Poincaré} \\ \text{qui citent un mathématicien et traitant de géométrie.} \end{array} \right.$$

4. Représentation des correspondants et des personnes citées

Lors de l'édition numérique du corpus, les acteurs ont tout d'abord été décrits assez brièvement par l'édition du nom, des prénoms, des dates de naissance et de décès, ainsi que par une courte description lorsque suffisamment d'information était disponible. Récemment, la volonté d'ajouter des détails pour les acteurs du corpus est apparue, notamment afin de mettre en avant leur place dans la chronologie poincaréienne, et dans l'optique de mettre en avant des réseaux de personnes, qu'ils soient déjà connus ou non. Ces réseaux peuvent correspondre à des réseaux privés, par le biais de relations familiales ou amicales ; à des réseaux scientifiques, par le biais de travaux communs, d'échanges scientifiques ; à des réseaux académiques, au travers des liens étroits qu'entretenait Poincaré avec certains enseignants et membres d'établissements de formation. Il est un enjeu majeur que de rendre compte des rôles tenus par les correspondants au sein de ce corpus.

Pour cela, les acteurs sont décrits par un ensemble de métadonnées précisant des informations à leur propos. Cela comprend tout d'abord un lieu de naissance indiquant la ville et le pays de naissance [14]. La ou les nationalités attachées aux personnes est indiquée en complément de ces informations.

Afin de rendre compte des activités des acteurs, un ensemble de classes correspondants à des situations sociales et professionnelles ont été créées. Ce travail est délicat, car il nécessite d'effectuer des simplifications de la réalité, qui peuvent parfois nous sembler inadaptées. Il peut néanmoins encourager la prise de recul et force parfois à se replonger dans leur histoire et à lever certaines ambiguïtés lorsque c'est possible. Plusieurs groupes principaux ont émergé, tels que l'ensemble des universitaires, des scientifiques, des ingénieurs, des académiciens, des

14. Se pose la problématique des évolutions administratives relatives aux pays. Pour certains acteurs, le pays de naissance n'existe plus en tant que tel. C'est notamment le cas pour Marie Skłodowska-Curie, née en 1869 à Varsovie dans le royaume du Congrès, qui était alors une province de l'empire Russe, et située géographiquement dans l'actuelle Pologne. Dans ces situations, le choix a été effectué de renseigner le pays de naissance tel qu'il existe actuellement. Cela est nécessaire afin de créer des relations entre le graphe du corpus de la correspondance et le graphe public GeoNames, où les pays sont listés selon les entités administratives associées telles qu'elles existent de nos jours. À chacune de ces entités sont associées des coordonnées géographiques.

éditeurs scientifiques, etc. Certains acteurs sont également associés à un rôle déterminé par leur profession principale, telle que la profession d'écrivain, de journaliste, d'avocat, de traducteur, etc. Il est important de souligner que la plupart des acteurs ne sont pas limités à une seule description, et que nombre d'entre eux appartiennent à plusieurs groupes.

De nombreux acteurs sont également indissociables de leurs rôles au sein d'administrations et de sociétés savantes. Il a été choisi de n'indiquer que les liens les plus significatifs, pouvant apporter un éclairage important aux échanges tenus avec Poincaré. Parmi les institutions les plus souvent citées, nous pouvons évoquer l'Académie des sciences, la *Royal Society*, le Cercle mathématique de Palerme, ou encore le Bureau des longitudes.

Pour certains acteurs, les échanges et les relations avec Poincaré sont dissociés de leurs fonctions ou des rôles qui leur sont généralement attribués en dehors du corpus. C'est pourquoi il est apparu nécessaire de distinguer des informations générales d'informations spécifiques au corpus. Dans ce contexte, un ensemble de propriétés ont été définies pour décrire la façon dont Poincaré connaît et interagit avec ces acteurs. Cinq types de réseaux ont été identifiés pour ce corpus : le réseau familial, le réseau amical, le réseau scientifique, le réseau mondain et le réseau polytechnicien. Encore une fois, de nombreux acteurs sont liés à plusieurs de ces réseaux. Par exemple, le mathématicien Paul Appell, camarade de Poincaré au lycée impérial de Nancy, collègue avec qui il a scientifiquement échangé tout au long de sa carrière qui était également un ami proche [15].

Mener cette édition numérique implique de devoir effectuer de nombreux choix pour représenter les données du corpus. Bien que ces choix puissent parfois être contraignants, car ils correspondent à une simplification de la réalité et ne traduisent que partiellement certains faits, cet exercice peut cependant encourager la prise de recul sur les objets étudiés et peut ainsi contribuer à renouveler les recherches. Dans ce contexte, il est important de rendre compte de ces choix pour éviter les ambiguïtés, pour renforcer la cohérence des travaux, et pour présenter la méthodologie de recherche utilisée qui pourrait inspirer d'autres chercheurs. À titre d'exemple, le tableau 1 regroupe des statistiques relatives aux cinq correspondants avec lesquels Poincaré a le plus échangé.

Dans la suite de cet article, nous présentons la façon dont les données du corpus peuvent être exploitées au travers d'outils numériques.

5. Un outil d'interrogation flexible pour explorer le corpus

5.1. L'idée de transformation de requêtes

Au sein du site dédié à Henri Poincaré, une interface utilisant l'outil Solr [19] permet d'effectuer des recherches plein texte pour les lettres ayant une transcription [16]. Par exemple, une recherche après la saisie du terme « géométrie » retourne un ensemble de 26 lettres rédigées ou reçues par Henri Poincaré. Bien qu'il soit utile dans de nombreux cas, cet outil ne profite pas des technologies du Web sémantique. Le langage SPARQL est plus expressif et permet de formuler des requêtes complexes. Cependant, l'utiliser nécessite de maîtriser une syntaxe particulière

15. Paul Appell a d'ailleurs publié en 1925 un ouvrage biographique afin de rendre hommage à son ami Henri Poincaré [18].

16. Environ 60% des lettres sont associées à une transcription rédigée en LaTeXpuis convertie en HTML.

Table 1
Statistiques sur les cinq correspondants avec lesquels Poincaré a le plus échangé.

Personne	Lettres échangées	Citations (lettre)	Citations (apparat)	Description
Mittag-Leffler, Gösta (1846-1927)	262	54	179	Mathématicien suédois, créateur d'*Acta Mathematica*
Guccia, Giovanni Battista (1855-1914)	72	7	34	Mathématicien italien, éditeur des *Rendiconti del Circolo Matematico di Palermo*
Darwin, George Howard (1845-1912)	50	2	5	Mathématicien et astronome britannique
Klein, Felix (1849-1925)	41	37	48	Mathématicien allemand
Saint-Arroman, Raoul de (1849-1915)	39	2	2	Écrivain, journaliste et haut fonctionnaire français

qui n'est pas facile d'utilisation pour tous les historiens et visiteurs du site. Dans ce contexte, des collaborations entre historiens et informaticiens s'intéressent à l'élaboration et aux usages d'outils de recherche interactifs pour explorer le corpus de la correspondance d'Henri Poincaré. En particulier, des travaux visent à proposer des outils basés sur un mécanisme d'interrogation flexible, s'intégrant dans le cadre du Web sémantique [20].

Imaginons un historien des sciences à la recherche d'informations concernant la *mécanique rationnelle* dans la correspondance d'Henri Poincaré. Un point de départ possible serait de s'intéresser aux échanges avec Paul Appell qui a rédigé plusieurs traités de mécanique rationnelle, notamment un premier volume paru en 1893 [21]. Soit $\mathcal{Q}$ une requête informelle [17] formulée par l'historien :

$$\mathcal{Q} = \left| \begin{array}{l} \text{Donner les lettres envoyées entre 1890 et} \\ \text{1895 par Paul Appell à Henri Poincaré et} \\ \text{qui mentionnent des travaux en mécanique.} \end{array} \right.$$

L'exécution de cette requête sur la base RDF du corpus de la correspondance pourrait retourner des résultats qui ne sont pas satisfaisants pour l'historien. Les raisons suivantes de cette insatisfaction peuvent être considérées :

- l'ensemble des résultats est trop grand : l'historien voudrait spécialiser sa requête ;
- l'ensemble des résultats est vide ou trop petit : il souhaiterait généraliser sa requête ;
- les résultats obtenus n'apportent pas de réponse à la problématique initiale : il devrait envisager de nouvelles pistes de recherche.

Une approche pour remédier à ce problème consiste en la définition et l'application de règles de transformation de requêtes. Ces règles peuvent être générales ou dépendantes d'un domaine. Par exemple, voici quatre nouvelles requêtes qui pourraient résulter de la transformation de $\mathcal{Q}$:

$$\mathcal{Q}_1 = \left| \begin{array}{l} \text{Donner les lettres envoyées entre 1890 et} \\ \text{1895 par Henri Poincaré à Paul Appell et} \\ \text{qui mentionnent des travaux en mécanique.} \end{array} \right.$$

17. Les requêtes utilisées dans ce document sont présentées de façon informelle dans un souci de lisibilité mais elles correspondent toutes à une requête SPARQL.

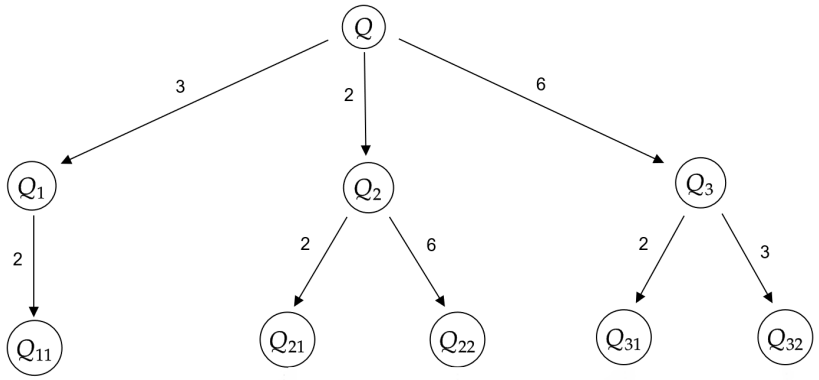


Figure 4 : Exemple d'un arbre de recherche tronqué à la profondeur 2.

$$\mathcal{Q}_2 = \left| \begin{array}{l} \text{Donner les lettres envoyées entre 1890 et} \\ \text{1895 par un mathématicien à Henri Poincaré et} \\ \text{qui mentionnent des travaux en mécanique.} \end{array} \right.$$

$$\mathcal{Q}_3 = \left| \begin{array}{l} \text{Donner les lettres envoyées entre 1890 et} \\ \text{1895 par Émile Picard à Henri Poincaré et} \\ \text{qui mentionnent des travaux en mécanique.} \end{array} \right.$$

$$\mathcal{Q}_4 = \left| \begin{array}{l} \text{Donner les lettres envoyées après 1895} \\ \text{par Paul Appell à Henri Poincaré et} \\ \text{qui mentionnent des travaux en mécanique.} \end{array} \right.$$

$\mathcal{Q}_1$ est générée en appliquant une règle d'échange de l'expéditeur et du destinataire de la lettre. $\mathcal{Q}_2$ est générée en appliquant une règle visant à remplacer une instance de classe par n'importe quel membre de cette classe. Dans notre cas, la ressource décrivant Paul Appell fait partie de la classe des mathématiciens. $\mathcal{Q}_3$ correspond à l'application d'une règle visant à remplacer l'un des correspondants par une personne avec laquelle il a collaboré. $\mathcal{Q}_4$ correspond à l'application d'une règle visant à modifier les bornes temporelles liées à la date d'écriture de la lettre. D'autres générations de requêtes peuvent être imaginées et sont dépendantes des règles existantes et des données de la base RDF. Il est possible de combiner plusieurs transformations de requêtes tant qu'un coût maximal n'a pas été atteint. En effet, dans le fonctionnement actuel de l'outil, un coût (défini comme un nombre positif) est associé à chaque règle de transformation de requêtes. Ainsi, l'application successive de règles de transformation correspond à l'exploration d'un arbre de recherche comme l'illustre la figure 4.

L'objectif de ce mécanisme de transformation est double. Tout d'abord, il permet de présenter des résultats sémantiquement proches de ceux correspondant aux critères de recherche et qui peuvent apporter une réponse au problème de recherche pour lequel la requête initiale avait été formulée. Mais ce mécanisme peut aussi faire émerger de nouvelles connaissances pour le domaine. En effet, dans le contexte de la correspondance d'Henri Poincaré, il peut permettre de

dégager de nouveaux liens ou d'affiner des liens existants entre des personnes, des institutions, etc.

Ce système d'interrogation flexible a fait l'objet de plusieurs publications, qui précisent notamment sa formalisation informatique en expliquant la syntaxe et l'algorithme permettant l'application des règles de transformation [20, 22, 23]. Nous proposons de nous intéresser à plusieurscas d'utilisation, en présentant des exemples issus du corpus de la correspondance d'Henri Poincaré.

5.2. Quelles règles de transformation pour quels usages ?

Certaines règles génériques, qui exploiteraient les relations entre les éléments de l'ontologie, peuvent présenter un intérêt pour proposer des alternatives lors des recherches, notamment par la généralisation de la requête initialement formulée. Le système SQTRL permet également de définir des règles dépendantes du domaine d'application, et peut ainsi encourager l'exploitation de connaissances spécifiques. Bien que les règles de transformation n'aient pas à être rédigées par les experts du domaine, il est important qu'ils participent activement à la réflexion sur leur formalisation et leurs possibles applications pour la recherche, ici, dans un cadre historique [18]. Afin d'exploiter les différentes données relatives aux acteurs, plusieurs règles de transformation ont été proposées. L'exploitation de ces informations permet d'offrir de nombreuses directions de recherche à un utilisateur explorant le corpus. Il est intéressant de noter qu'au-delà de l'étude des travaux de Poincaré, le corpus peut éclairer les travaux scientifiques de certains acteurs ayant interagi avec Poincaré. Dans certains cas, les correspondances avec des acteurs sont conséquentes et fournissent une base de travail intéressante. Dans d'autres cas, cet ensemble de lettres peut n'avoir été que partiellement constitué, ou être inexistant, ou bien ne pas contenir les informations pouvant apporter des éléments de réponse aux problématiques de recherche. Cependant, il peut exister un nombre important de lettres mentionnant ce correspondant, et au sein desquelles des éléments historiques et scientifiques pertinents peuvent éclairer le questionnement initial. Dans ce contexte, une règle est proposée pour remplacer un correspondant en tant que personne citée dans le corps de la lettre dans une requête SPARQL. Il est également intéressant de proposer la fonctionnalité inverse qui remplace une personne citée pour qu'elle soit un correspondant des lettres recherchées. Il existe également des versions alternatives de ces règles qui s'appuient sur la propriété utilisée pour indiquer les personnes citées dans l'apparat critique.

Imaginons un chercheur s'intéressant aux travaux du mathématicien allemand Karl Weierstrass (1815-1897). Ce dernier, lauréat de la médaille *Copley* de la *Royal Society* de Londres en 1897, a proposé des travaux de premier plan en analyse qu'il a appliqués pour des contributions significatives dans le champ du calcul des variations [24]. Au cours de sa carrière, il a eu pour disciples de nombreux mathématiciens avec lesquels Henri Poincaré a échangé au cours de sa vie parmi lesquels nous pouvons citer Georg Cantor, Lazarus Fuchs et Sofia Kovalevskaïa. Au sein du corpus, il existe uniquement 4 lettres avec Karl Weierstrass défini en tant que correspondant. Cependant, nous remarquons qu'il existe des lettres où ce mathématicien est

18. En effet, l'idée du système SQTRL est notamment d'encourager la collaboration entre des chercheurs spécialistes d'un domaine, et des informaticiens qui seraient à l'aise avec la syntaxe des règles et le cadre technologique du Web sémantique.

explicitement cité dans le corps (49 lettres) ou dans l'apparat critique (64 lettres), principalement dans des échanges avec Gösta Mittag-Leffler. Dans ce genre de situations, une application de règle pourrait transformer la requête informelle $\mathcal{Q}$ en $\mathcal{Q}'$.

$$\mathcal{Q} = \left| \begin{array}{l} \text{Donner les lettres échangés entre} \\ \text{Henri Poincaré et Karl Weierstrass} \\ \text{qui traitent d'analyse.} \end{array} \right. \overset{\text{subsCorr}}{\rightsquigarrow} \mathcal{Q}' = \left| \begin{array}{l} \text{Donner les lettres dont Henri} \\ \text{Poincaré est correspondant citant} \\ \text{Karl Weierstrass et qui traitent d'analyse.} \end{array} \right.$$

Dans une autre démarche, plusieurs règles ont été définies, suivant le même motif, afin d'exploiter les métadonnées décrivant les acteurs. Si un acteur est explicitement mentionné dans le corps d'une requête SPARQL, il est possible de généraliser cette requête en s'appuyant sur une de ses caractéristiques. Par exemple, la requête suivante recherche les lettres échangées avec Giovanni Battista Guccia :

$$\mathcal{Q} = \left| \begin{array}{l} \text{Donner les lettres échangées avec Giovanni Battista Guccia.} \end{array} \right.$$

À partir des informations le concernant, le système peut proposer plusieurs transformations dont celles générant les quatre requêtes suivantes :

$$\mathcal{Q}_1 = \left| \begin{array}{l} \text{Donner les lettres échangées avec} \\ \text{un mathématicien.} \end{array} \right. \qquad \mathcal{Q}_2 = \left| \begin{array}{l} \text{Donner les lettres échangées avec} \\ \text{un italien.} \end{array} \right.$$

$$\mathcal{Q}_3 = \left| \begin{array}{l} \text{Donner les lettres échangées avec un} \\ \text{membre du cercle mathématique de Palerme.} \end{array} \right. \qquad \mathcal{Q}_4 = \left| \begin{array}{l} \text{Donner les lettres échangées avec un} \\ \text{éditeur scientifique.} \end{array} \right.$$

Des motifs de co-occurrences pourraient appuyer des transformations de requêtes lors de recherches sur le corpus. Par exemple, une possibilité serait de s'intéresser aux liens entre les correspondants et les thèmes associés aux lettres. Lors d'une recherche impliquant un correspondant, il peut être pertinent de proposer de rechercher des lettres avec un autre correspondant, qui a échangé avec Poincaré sur un même thème.

Une autre forme de transformation pourrait s'appuyer sur des liens entre une période temporelle et un thème. Partant d'un thème indiqué comme critère d'une recherche sur la correspondance, il pourrait être intéressant de proposer une nouvelle requête avec des critères de bornes temporelles construits à partir de la période temporelle où apparaît ce thème. Par exemple, la requête suivante s'intéresse au thème « fonction fuchsienne ».

$$\mathcal{Q} = \left| \begin{array}{l} \text{Donner les lettres traitant de fonctions fuchsiennes.} \end{array} \right.$$

Au sein de la correspondance, 12 lettres, rédigées entre avril 1881 et décembre 1882, traitent des fonctions fuchsiennes. En appliquant la transformation évoquée plus haut, la requête deviendrait :

$$\mathcal{Q}' = \left| \begin{array}{l} \text{Donner les lettres rédigées entre le 11 avril 1881 et le 05 décembre 1882.} \end{array} \right.$$

6. Des systèmes pour l'exploration du corps

Dans la première partie de ce chapitre, nous avons évoqué plusieurs règles de transformations et leurs possibles applications pour aider à mener des recherches historiques sur le corpus de la

correspondance d'Henri Poincaré. Nous proposons de décrire deux interfaces de recherches qui s'appuient sur l'utilisation de ce mécanisme. La première est un système de recherche s'appuyant sur des formulaires. Le deuxième est un outil de navigation qui exploite les similarités entre des ressources d'un graphe RDF. D'autres outils, qui ne s'appuient pas sur le mécanisme d'interrogation flexible seront également brièvement décrits.

6.1. Interface fondée sur des formulaires

Une première application du mécanisme d'interrogation flexible se retrouve dans un système de recherche simple, qui permet de générer des requêtes SPARQL via une interface fondée sur des formulaires. Le formulaire affiché s'adapte au type de ressources recherchées (articles, documents, lettres, personnes), plusieurs champs sont proposés à la saisie. Pour plusieurs d'entre eux, un mécanisme d'autocomplétion est utilisé, qui s'appuie sur l'utilisation des étiquettes (*label*) associés aux identifiants de la base. Par exemple, dans la base du corpus, Henri Poincaré est associé à l'identifiant `henripoincare.fr/api/items/843` mais l'interface affiche la chaîne "Poincaré, Henri (1854-1912)". À tout moment, l'utilisateur peut visualiser dans un champ dédié la requête SPARQL qui est générée. Une fois une requête exécutée, les résultats correspondants sont affichés dans un tableau et peuvent être exportés au format CSV. Le système propose également des requêtes alternatives s'appuyant sur l'application de règles de transformation. Sélectionner une transformation va permettre de générer une nouvelle. À tout moment, l'utilisateur peut visualiser la requête initialement formulée ainsi que la requête générée après application d'une règle de transformation. Les différents résultats sont agrégés dans le même tableau, en distinguant les résultats additionnels des résultats initiaux par l'utilisation de couleurs. Cet outil permet à un utilisateur de formuler des requêtes SPARQL sans aucune connaissance technique tout en permettant l'utilisation du système d'interrogation flexible.

6.2. Outil de navigation

Une autre forme d'exploration du corpus est relative à la notion de navigation qui part du constat qu'il est fréquent de débuter par l'étude d'une ressource du corpus avant de s'intéresser à d'autres ressources liées ou similaires. Dans ce contexte, un outil de navigation a été développé pour ce corpus. Le principe est de partir d'une lettre du corpus et de retrouver des ressources présentant des similarités. Pour cela, après la sélection de la ressource initiale, le système propose automatiquement plusieurs conditions de recherche que l'utilisateur peut sélectionner comme critère de recherche (positivement ou négativement). Pour l'exemple présenté en figure 5, cela conduit à la génération de sept conditions relatives à l'expéditeur, au destinataire, aux thèmes et aux personnes citées. À côté de chacune d'entre elles, il peut également retrouver un entier indiquant le nombre de ressources correspondant à ce critère dans le graphe. Après sélection des conditions qui l'intéressent, il a la possibilité de générer et d'exécuter une requête SPARQL. Les résultats apparaissent alors dans la partie droite de l'interface, avec des détails pour chaque lettre retrouvée. Il est possible de filtrer les résultats selon la date de rédaction des lettres, de manière à s'intéresser à une période temporelle spécifique. Des fonctionnalités annexes comme l'affichage d'un graphique correspondant à la distribution temporelle des résultats et l'export des données au format CSV sont également proposés. À partir des résultats, l'utilisateur peut

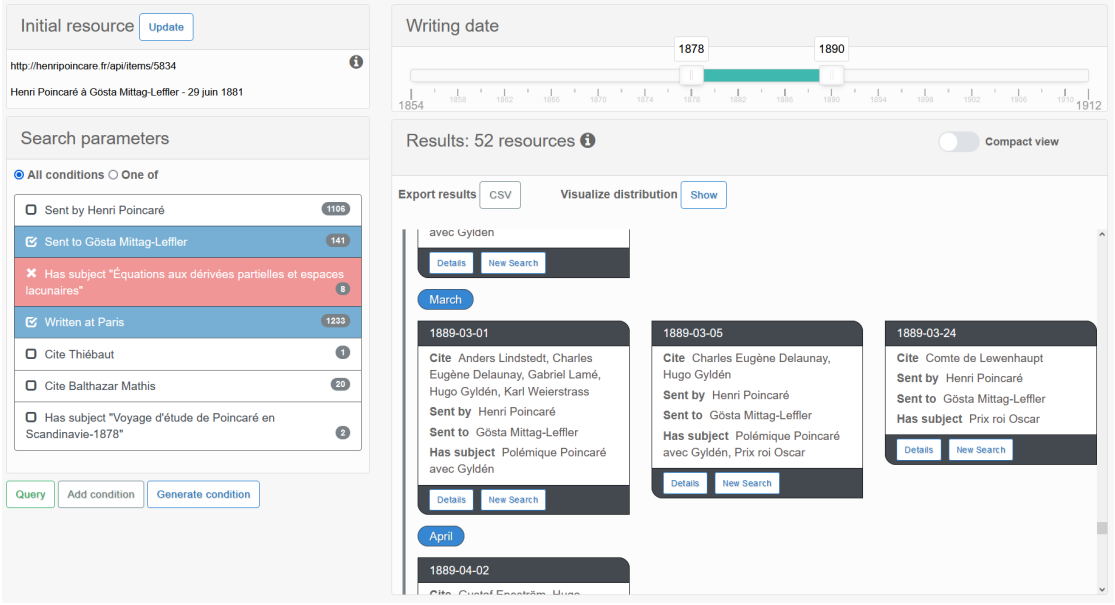


Figure 5 : L'outil de navigation en utilisation pour explorer le corpus de la correspondance d'Henri Poincaré.

sélectionner une nouvelle ressource qui sera le point de départ d'une nouvelle recherche. Les conditions de recherche seront alors mises à jour. Il est ainsi possible de naviguer de ressource en ressource.

Pour aller plus loin, le système propose la génération automatique de conditions additionnelles en appliquant des règles de transformation de requêtes. En effet, dans certaines situations, les conditions générées peuvent ne pas être suffisantes pour fournir des résultats intéressants ou surprenants.

Ce système peut être intéressant pour explorer les données d'une façon originale. Dans certaines situations, la plupart des ressources liées pourraient être déjà connues des historiens, mais dans d'autres cas, l'outil pourrait mettre en évidence des liens inattendus entre les ressources et être le point de départ de nouvelles considérations historiques. À titre d'exemple, ces nouveaux liens peuvent révéler certains motifs entre des thèmes scientifiques et des individus (en tant que correspondants ou personnes citées), en particulier pour les acteurs considérés comme mineurs et pour lesquels aucun travail de recherche n'a été initié.

Une vidéo de démonstration de cet outil est disponible en ligne [19] et le code source est disponible sur un dépôt GitHub public [20]. Cet outil a été pensé dès le départ, comme un système réutilisable et est donc accompagné d'un fichier de configuration qui permet de se connecter à d'autres points d'accès SPARQL. Les différents éléments présentés dans l'interface sont tous paramétrables. Une documentation relative à son installation et sa configuration, et la présentation d'essais avec des données de DBpedia sont notamment fournies sur le dépôt GitHub.

19. https://videos.ahp-numerique.fr/videos/watch/f90ff003-39db-4b4c-ade6-fb18b86d9244.
20. https://github.com/nlasolle/rdf_navigation_tool

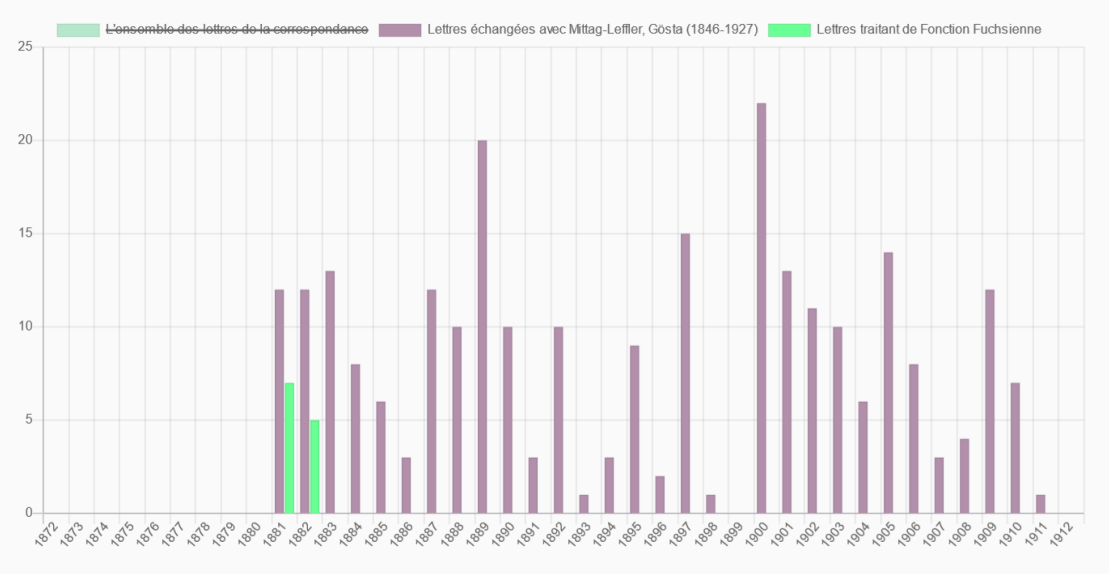


Figure 6 : Exemple d'affichages de distributions relatives à la rédaction de lettres respectant certains critères.

6.3. Des outils s'appuyant sur des statistiques et des données spatio-temporelles

D'autres outils de recherche ont été mis en place et sont à la fois à destination du grand public souhaitant découvrir le corpus et à destination des historiens souhaitant mener des études quantitatives. Le code source de ces différents outils est accessible sur un dépôt GitHub public [21] et une vidéo de démonstration est disponible en ligne [22]. Contrairement à l'outil de navigation, il n'y a pas de système de configuration qui permet de se connecter à un autre corpus mais il est cependant possible de s'inspirer des outils proposés en réutilisant certaines parties de l'application.

Le premier outil permet à un utilisateur de sélectionner un correspondant ou un thème, et de visualiser la distribution des lettres associées. Pour cela, l'outil s'appuie sur l'année de rédaction des lettres. Il permet notamment de superposer différentes distributions, ce qui peut être utile pour rapidement mettre en évidence des motifs ou pour valider certaines hypothèses. Le graphique peut être exporté en tant qu'image et il est également possible d'exporter les données correspondantes au format CSV. Par exemple, la figure 6 superpose la distribution des lettres échangées entre Henri Poincaré et Thomas Craig et la distribution des lettres ayant pour thème la géométrie.

Le deuxième outil correspond à un simple tableau dynamique, qui regroupe des statistiques relatives aux personnes du corpus. Pour chacune d'entre elles, le nombre de lettres échangées avec Poincaré, le nombre de lettres la citant, et le nombre de lettres dont l'apparat la cite est indiqué. En cliquant sur une ligne du tableau, il est possible d'accéder à la page du site qui donne

21. https://github.com/nlasolle/ahpo-data-exploration-tools
22. https://videos.ahp-numerique.fr/w/gjj2DJ9mZmVNKehwuDgWFk

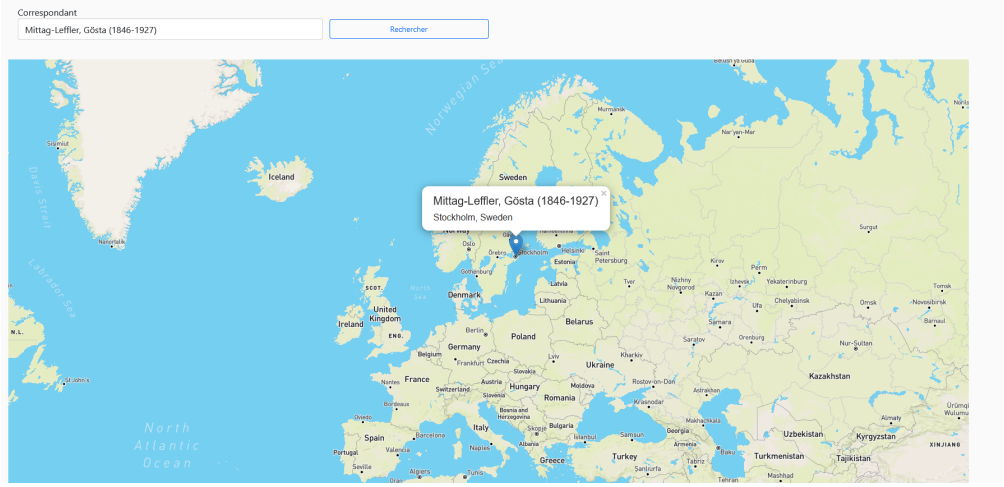


Figure 7 : Outil de visualisation des lieux de naissance associés aux personnes du corpus.

des détails sur la personne [23]. Ce tableau s'appuie sur l'interrogation du point d'accès SPARQL qui permet une mise à jour continue des données présentées.

Enfin, le dernier outil exploite les données géographiques relatives aux personnes du corpus. Le premier correspond à une interface où un utilisateur peut visualiser des lieux de naissance, sous la forme de marqueurs par rapport à une personne choisie (voir figure 7, p. 19). Par exemple, en sélectionnant Felix Klein, la carte met en évidence son lieu de naissance (Düsseldorf, Allemagne) et les villes d'exercice connues (plusieurs villes allemandes). Une autre carte propose d'afficher les lieux de naissance connus de l'ensemble des correspondants avec lesquels Poincaré a échangé.

7. Conclusion

Le corpus de la correspondance d'Henri Poincaré est un corpus riche, contenant de nombreuses informations pour des travaux relevant de l'histoire des sciences et de l'histoire sociale. L'usage d'outils numériques peut appuyer des recherches historiques s'y intéressant et encourager la découverte de liens. Un constat porté par l'équipe est que l'édition numérique du corpus est un travail conséquent qui nécessite une phase de réflexion préliminaire visant à définir les données d'intérêt, la structure de la base, les choix ontologiques. Il est nécessaire de déterminer les informations qui devront apparaître dans la base, quelle sémantique associer à chaque métadonnée, comment structurer la base, quelles ontologies réutiliser et sous quelles formes. Il y a ensuite le travail d'alimentation de la base qui consiste à éditer les faits sous la forme de triplets. Cela nécessite de s'accorder quant aux sources utilisées pour décrire les éléments du corpus et de déterminer jusqu'où aller dans les descriptions. Il peut être tentant d'être le plus exhaustif possible, mais cela peut parfois conduire à des approximations, ou à intégrer des données qui

23. Par exemple, la page http://henripoincare.fr/s/correspondance/item/346 offre une synthèse des informations relatives au mathématicien allemand Felix Klein.

ne présentent qu'un intérêt mineur pour la recherche historique tout en nécessitant un travail conséquent pour retrouver les informations. Les contraintes des projets de recherche font que des tâches relatives à l'édition sont parfois limitées dans le temps, et il faut donc effectuer des choix en conséquence pour garantir un corpus numérique cohérent bien qu'incomplet.

Le mécanisme d'interrogation flexible peut présenter plusieurs intérêts dans le cadre de l'étude. Il peut tout d'abord permettre d'accéder plus rapidement à des informations utiles. Pour les personnes ne connaissant pas le corpus, cela peut être l'occasion d'être guidé dans leurs recherches pour comprendre son organisation. Ce système peut également s'inscrire dans une méthode heuristique en permettant de mettre en avant des liens qui, bien souvent pourraient être déjà connus mais dans certains cas être inattendus. Dans le cadre de l'étude d'un corpus historique tel que celui de la correspondance d'Henri Poincaré, il peut notamment mettre en avant des liens avec des acteurs parfois qualifiés de secondaires mais qui peuvent présenter un intérêt lors de l'étude de la vie et de l'œuvre de Poincaré, ou lors de l'étude d'institutions savantes de la fin du XIX^{e} siècle et du début du XX^{e} siècle.

Remerciements

Ce travail a bénéficié d'une aide de l'État, gérée par l'Agence Nationale de la Recherche, au titre du projet Investissements d'Avenir Lorraine Université d'Excellence, portant la référence ANR-15-IDEX-04-LUE.

Un grand merci à toutes les personnes ayant participé, de près ou de loin, à ce projet d'édition et d'exploitation numérique de la correspondance d'Henri Poincaré.

Références

[1] P. Nabonnand (Ed.), La correspondance entre Henri Poincaré et Gösta Mittag-Leffler, Birkhäuser, 1998.

[2] S. Walter, É. Bolmont, A. Coret (Eds.), La correspondance entre Henri Poincaré et les physiciens, chimistes et ingénieurs, Birkhäuser, 2007. doi :10.1007/978-3-7643-8303-9.

[3] S. Walter, P. Nabonnand, R. Krömer, M. Schiavon (Eds.), La correspondance entre Henri Poincaré, les astronomes, et les géodésiens, Birkhäuser, 2016. doi :10.1007/978-3-7643-8293-3.

[4] L. Rollet (Ed.), La correspondance de jeunesse d'Henri Poincaré : les années de formation. De l'École polytechnique à l'École des Mines (1873-1878), Publications of the Henri Poincaré Archives, Birkhäuser, 2017. doi :10.1007/978-3-319-55959-9.

[5] L. Rollet, P. Nabonnand, Éditer la correspondance d'Henri Poincaré, in : F. Henryot (Ed.), L'historien face au manuscrit. Du parchemin à la bibliothèque numérique, UCL-Presses Universitaires de Louvain, 2012, pp. 285–304. URL : http://books.openedition.org/pucl/1282.

[6] J. Peiffer, Faire des mathématiques par lettres, Revue d'histoire des mathématiques 4 (1998) 143–157.

[7] A. Gilroy, W. M. Verhoeven, Epistolary histories : letters, fiction, culture, University of Virginia Press, 2000.

[8] C. D. Conner, Histoire populaire des sciences, Editions l'Echappée, Paris, 2011.

[9] H. Zinn, Une histoire populaire des États-Unis. De 1492 à nos jours, Agone, Paris, 2002.
[10] G. Noiriel, Une histoire populaire de la France : De la guerre de Cent Ans à nos jours, Agone, Paris, 2018.
[11] A. Boutroux, Vingt ans de ma vie, simple vérité : la jeunesse d'Henri Poincaré racontée par sa sœur (1854-1878), Histoire des sciences, Hermann, 2012.
[12] L. Rollet, Jeanne louise poulain d'andecy, épouse poincaré (1857-1934), Bulletin de la Société des amis de la Bibliothèque et de l'Histoire de l'École polytechnique (2012) 18–27.
[13] C. Boulaire, R. Carabelli, Du digital naive au bricoleur numérique : les images et le logiciel Omeka, in : E. Cavalié, F. Clavert, O. Legendre, D. Martin (Eds.), Expérimenter les humanités numériques. Des outils individuels aux projets collectifs, Les Presses de l'Université de Montréal, Montréal, Québec, 2017, pp. 81–103. URL : http://www.parcoursnumeriques-pum.ca/du-digital-naive-au-bricoleur-numerique.
[14] N. Larrousse, J. Marchand, A Techno-Human Mesh for Humanities in France : Dealing with preservation complexity, in : Pierazzo, Elena and Ciotti, Fabio (Ed.), DH 2019, Utrecht, Netherlands, 2019. URL : https://hal.archives-ouvertes.fr/hal-02153016.
[15] F. Manola, E. Miller, B. McBride, et al., RDF 1.1 Primer, 2014. URL : https://www.w3.org/TR/rdf-primer, dernière consultation : juillet 2021.
[16] N. Lasolle, P. Willaime, Sémantiser Omeka S : pourquoi et comment ?, in : Omeka - Projets scientifiques, culturels et/ou documentaires, Nancy, France, 2020. URL : https://hal.univ-lorraine.fr/hal-02973316.
[17] S. Harris, A. Seaborne, E. Prud'hommeaux, SPARQL 1.1 Query Language, recommandation du W3C, 2013. URL : https://www.w3.org/TR/sparql11-query/, dernière consultation : juillet 2021.
[18] P. Appell, Henri Poincaré, Nobles vies - Grandes œuvres, Plon, 1925.
[19] T. Grainger, T. Potter, Solr in action, Manning Publications Co., 2014.
[20] O. Bruneau, E. Gaillard, N. Lasolle, J. Lieber, E. Nauer, J. Reynaud, A SPARQL Query Transformation Rule Language — Application to Retrieval and Adaptation in Case-Based Reasoning, in : D. Aha, J. Lieber (Eds.), Case-Based Reasoning Research and Development. ICCBR 2017, Lecture Notes in Computer Science, Springer, Cham, 2017, pp. 76–91.
[21] P. Appell, Traité de mécanique rationnelle, volume 1, Gauthier-Villars, 1893.
[22] O. Bruneau, N. Lasolle, J. Lieber, E. Nauer, S. Pavlova, L. Rollet, Applying and Developing Semantic Web Technologies for Exploiting a Corpus in History of Science : the Case Study of the Henri Poincaré Correspondence, Semantic Web – Interoperability, Usability, Applicability (2021) 359–378. doi :10.3233/SW-200400.
[23] N. Lasolle, O. Bruneau, J. Lieber, E. Nauer, S. Pavlova, Assisting the RDF Annotation of a Digital Humanities Corpus Using Case-Based Reasoning, in : J. Z. Pan, V. Tamma, C. d'Amato, K. Janowicz, B. Fu, A. Polleres, O. Seneviratne, L. Kagal (Eds.), The Semantic Web - ISWC 2020, Springer, Cham, 2020, pp. 617–633.
[24] P. Dugac, Eléments d'analyse de Karl Weierstrass, Archive for History of Exact Sciences 10 (1973) 41–176. URL : http://www.jstor.org/stable/41133363.

Les publications scientifiques en archéologie au format électronique : du CD-Rom au Web sémantique

Scientific publications in archaeology using digital format : from CD-Rom to the Semantic Web

Olivier Marlet[1,*], Xavier Rodier[1]

[1]*UMR7324 CITERES-LAT, CNRS, Université de Tours, France*

Abstract

For several years, we have been experimenting with new formats for the scientific publication of archaeological excavation results. In 2014, the results of the excavations on the site of the Château de Tours (Indre-et-Loire, France) were published in a mixed paper/digital format. In 2021, the publication of the excavations at the Marmoutier Abbey in Tours is entirely electronic in the form of a website. How then to retranscribe in electronic format what makes the strength of the paper reading ? The logicist publication of the excavations of the parish center of Rigny (Indre-et-Loire, France) in 2020 also marks an important turning point since the structure of the publication itself is new for an archaeological monograph. Each of these publications has taken advantage of the digital technologies available to offer the scientific community other ways of using and structuring information. In the age of the semantic web, these publication projects are part of the process of FAIRizing data for Open Science.

Keywords

scientific publishing, archaeology, logicism, Semantic Web, FAIR principles

1. Introduction

Bien avant l'apparition du concept d'Humanité Numérique, la recherche archéologique a exploité les outils informatiques disponibles, tant pour la gestion et l'analyse que pour la communication des résultats. Avec l'arrivée d'Internet, ont fleuri de nombreux sites Web pour présenter les données et les résultats scientifiques. Le Laboratoire Archéologie et Territoires (UMR 7324 CITERES-LAT) a exploré ainsi plusieurs pistes depuis 1998 pour moderniser les publications scientifiques et profiter des avantages qu'offre le support numérique. Cette réflexion tient compte de l'évolution technique des solutions numériques disponibles, en particulier pour la publication des données en archéologie. Les choix effectués pour les projets présentés dans cet article sont en fonction de l'état des connaissances techniques au moment de la préparation de ces publications.

Parallèlement aux questions d'éditions numériques, CITERES-LAT a participé activement aux travaux du consortium MASA (Mémoire des Archéologues et des Sites Archéologiques) dont l'objectif principal est d'accompagner les archéologues vers les bonnes pratiques numériques.

. *Workshop on Digital Humanities and Semantic Web*

∗. Corresponding author.

. olivier.marlet@univ-tours.fr (O. Marlet); xavier.rodier@univ-tours.fr (X. Rodier)

. https://www.univ-tours.fr/annuaire/m-olivier-marlet (O. Marlet); https://www.univ-tours.fr/annuaire/m-xavier-rodier (X. Rodier)

. 0000-0001-9422-1418 (O. Marlet); 0000-0002-1243-3167 (X. Rodier)

Depuis 2016, CITERES-LAT s'est investi dans une approche qualitative des données numériques selon les principes FAIR [1] (Faciles à trouver, Accessibles, Interopérables et Réutilisables). Ce processus s'inscrit dans la longue durée, surtout lorsqu'il s'agit de corpus constitués avant que les Principes FAIR ou même les « *5 Stars Linked OpenData* » [2] n'entrent en scène dans le monde de la recherche. Pour les publications, la mise en œuvre des bonnes pratiques implique, comme pour les données, un travail conséquent.

À partir de trois exemples de publications de CITERES-LAT, cet article propose d'offrir une vision critique des avantages et des inconvénients des solutions retenues selon les techniques disponibles au moment de leur réalisation. Nous présenterons donc à la fois les premières réflexions sur le CD-Rom jusqu'à des solutions fondées sur les principes de publication logiciste de Jean-Claude Gardin [1].

2. Du CR-Rom au site Web

2.1. La fouille du château de Tours

La fouille du château de Tours a été conduite de 1974 à 1978 par Henri Galinié. Au-delà des rapports scientifiques annuels et après le temps long de l'exploitation des données, le travail de publication a commencé en 1998. Le processus d'informatisation des données du laboratoire était engagé depuis 1990, alors que les données de cette fouille n'existaient qu'au format papier. Le choix s'est porté sur une publication mixte, associant un ouvrage imprimé à un CD-Rom. Ce CD-Rom ne devait pas être un doublon de l'ouvrage imprimé mais comporter plusieurs parties pour enrichir le texte de synthèse de l'ouvrage papier.

Cette partie numérique donne accès à l'enregistrement de terrain, à des analyses de collections (céramique, verre) et à une visite du site à travers quelques illustrations commentées des vestiges et des objets découverts (Fig. 1). L'intégralité des donnés a été numérisée en se limitant toutefois à la stricte granularité d'information nécessaire à l'articulation avec les démonstrations de la synthèse. Cela a démontré, comme l'indique Henri Galinié dans la préface de la publication que « l'informatisation systématique des données a alors mis au jour de façon catégorique des contradictions que le traitement traditionnel avait laissé esquiver ou sous-estimer ». Les révisions nécessaires pour lever ces contradictions expliquent en grande partie que le travail sur cette publication ait pu courir sur plus de dix ans, étant donné la quantité considérable de données produites par cette fouille.

Le choix du CD-Rom, encore pertinent en 1998, s'est révélé inapproprié quand la publication est devenue imminente en 2012. Il a alors été décidé de convertir la partie électronique de la publication en un site Web. Le CD-Rom avait été structuré depuis 1998 selon une arborescence de fichiers HTML statiques dans lesquels la navigation s'effectue via des liens hypertextes. Un choix s'est alors offert à nous : soit conserver cette architecture statique de fichiers HTML tout en profitant de l'informatisation des données pour produire une extraction de la base de données, soit élaborer un site Web dynamique alimenté par une base de données relationnelles.

1. En 2016, le groupe FORCE11 publie *FAIR Guiding Principles for scientific data management and stewardship* dans la revue *Scientific Data* : https://www.nature.com/articles/sdata201618.

2. En 2010, Tim Berners-Lee, un des inventeurs du Web, a proposé une méthode basée sur 5 étoiles pour évaluer le niveau d'ouverture des données sur le « Linked Open Data » : https://5stardata.info/fr/.

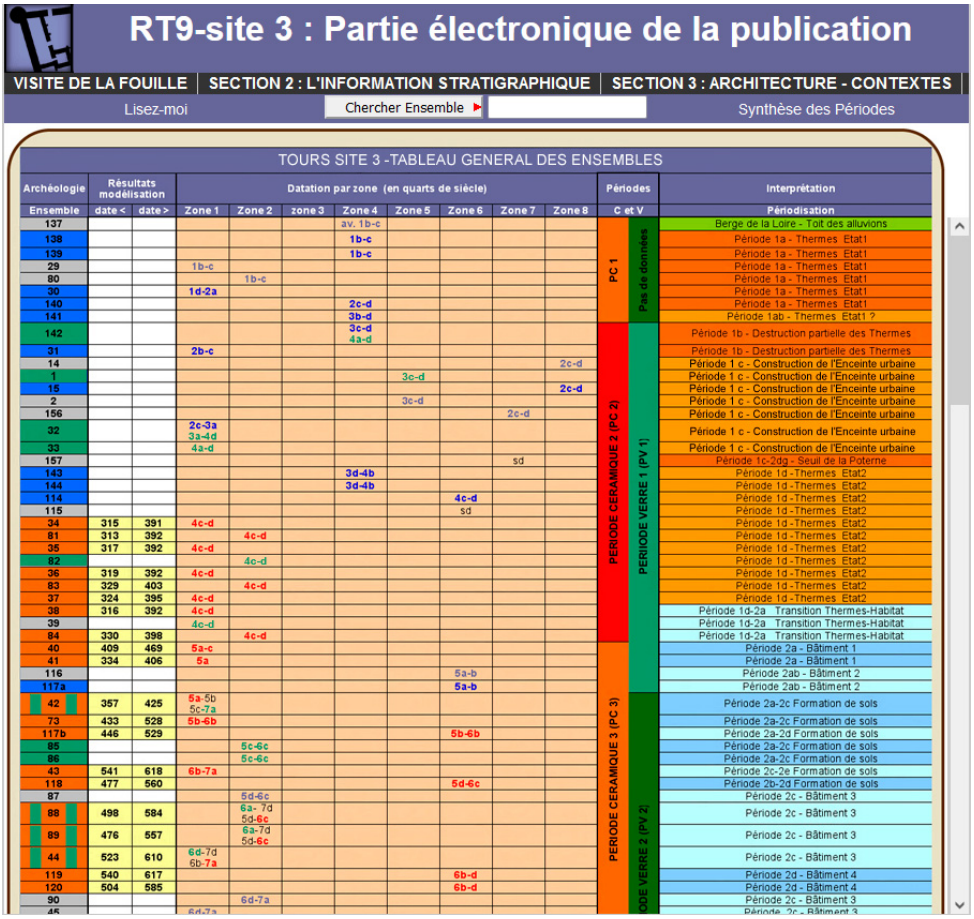


Figure 1: Partie électronique de la publication sur le site 3 à Tours.

Malheureusement, le manque de moyens nous a conduit à en rester à la reproduction de la structure du CD-Rom directement sur un site Web. Cela posait bien entendu le problème des mises à jours extrêmement laborieuses, mais cette question était secondaire puisque les données mises à disposition correspondent à l'état des connaissances publié. Des mises à jour seraient d'ailleurs potentiellement incohérentes avec la synthèse du volume imprimé. La publication a paru en 2014 [2], sous la forme d'un volume de synthèse imprimé, accompagné d'un site Web constitué de pages HTML statiques [3] comprenant une visite de la fouille et deux sections intitulées « L'information stratigraphique » et « Architecture – Contextes ».

D'un point de vue des bonnes pratiques numériques, il est évident que cette partie électronique ne répond pas du tout aux recommandations déjà exprimées à l'époque par les « 5 Stars Linked Open Data » de Tim Berners-Lee (les principes FAIR n'ont été publiés que deux ans après la publication). Il ne s'agit pas là de mauvais choix mais simplement d'une situation particulière justifiée par une mise en œuvre inscrite dans la longue durée qui a rendu désuet les choix initiaux.

3. http://citeres.univ-tours.fr/rt9/.

L'accumulation de pages statiques (plus de 3000 fichiers HTML) constituent un frein évident à la mise en place du site Web, à sa mise à jour et à la migration éventuelle vers d'autres supports. Sauf si l'obsolescence du site Web dans le futur nous l'imposait, ces deux derniers points sont *a priori* écartés pour une publication finale. De plus, la structuration de la partie électronique élaborée pour cette publication n'est pas reproductible automatiquement. Toutefois, d'un point de vue strictement scientifique, les sections numériques sont parfaitement fonctionnelles et offrent un accès structuré aux corpus de preuves structurés, issus de l'enregistrement de terrain et des analyses des collections d'artefacts.

2.2. Le CD-Rom en complément d'un volume imprimé

D'autres publications de CITERES-LAT ont utilisé le CD-Rom en complément d'un volume imprimé. Tout en tirant partie des apports du numérique (vidéos, navigation), ils ne présentent que des annexes à une publication traditionnelle et n'ont pas ou peu conduit à en repenser fondamentalement la structure. C'est le cas de *Tours antique et médiéval, 40 ans d'archéologie urbaine* [3], dont le CD-Rom présente une visite en image d'une exposition, un PDF des sources et références bibliographiques et plusieurs films courts présentant des restitutions 3D de différents monuments de la ville. En 2016, le contenu de ce CD-Rom a été mis en ligne [4] pour en faciliter à la fois l'accessibilité et la conservation.

Les publications du Projet Collectif de Recherche sur la céramique médiévale et moderne du Centre-Ouest de la France ont également fait appel au CD-Rom. La première a paru en 2003 accompagnée d'un CD-Rom comprenant 700 fichiers HTML statiques qui fournissent de manière structurée les sources typo-chronologiques sur lesquels s'appuient la publication [4]. La seconde a paru en 2013, elle-aussi accompagnée d'un CD-Rom comprenant 3700 fichiers HTML statiques, toujours parfaitement structuré et plus élaboré que le précédent [5]. Lors de la publication de 2013, le contenu du CD-Rom a également été mis en ligne [5]. Le même traitement a été appliqué à celui de 2003 [6]. Dans les deux cas, le format des fichiers et le contenu est rigoureusement identique sur le CD-Rom et sur le site Web. La prochaine publication de ce projet collectif de recherche est prévue pour 2022 et sera exclusivement en ligne [7] selon le format logiciste de Jean-Claude Gardin, suivant l'exemple de la publication de Rigny (voir *infra*, section 4), changeant complètement les modalités d'écriture et de lecture.

3. La publication Web des fouilles de l'hôtellerie de Marmoutier

3.1. Lecture et navigation

Au-delà du débat sur le plaisir de lire un support imprimé plutôt que sur écran, la publication Web offre des possibilités de lecture inenvisageable avec le support imprimé.

Ainsi, la publication Web des fouilles de l'hôtellerie de Marmoutier [8] [6] propose deux ni-

4. http://citeres.univ-tours.fr/tam/.
5. http://citeres.univ-tours.fr/cera2013/.
6. http://citeres.univ-tours.fr/cera2003/.
7. https://ceramedvaldeloire.huma-num.fr/.
8. https://marmoutier.univ-tours.fr/.

veaux de lecture : le premier synthétique donnant les grandes lignes de l'argumentation de l'archéologue, le second détaillé donnant accès au raisonnement et aux preuves sur lesquelles s'appuie l'argumentation. Le lecteur peut ainsi profiter d'une lecture rapide, constituée d'un texte synthétique et d'une sélection de figures pertinentes à ce niveau. Il s'agit alors d'une lecture proche d'une simple consultation comme on pourrait le faire d'un ouvrage en le feuilletant pour se faire une idée de son contenu. À tout moment, le lecteur peut basculer vers le texte détaillé pour accéder à l'argumentaire complet et aux corpus de preuves archéologiques (fiches d'enregistrement des couches, mobilier, etc.) dans la base de données en ligne ArSol [9] (Archives du Sol, [7]) qui permet la gestion des archives de fouilles du LAT. Donner l'accès aux preuves, c'est-à-dire les sources d'information mobilisées telles qu'elles ont été enregistrées sur le terrain, permet de donner plus de légitimité à l'argumentation proposée [8].

Un autre avantage de la publication Web est incontestablement l'exploitation des liens hypertextes. D'un simple clic depuis un appel dans le texte, il est possible d'afficher en surimpression à l'écran des informations complémentaires telle qu'une illustration, une référence bibliographique complète, l'équivalent d'une note de bas de page, un lien vers une ressources extérieures (les preuves dans ArSol par exemple).

Pour la publication des fouilles de l'hôtellerie de Marmoutier, la gestion des renvois a également été améliorée et ne se limite plus à de simples liens hypertextes. En effet, renvoyer à une autre section du texte implique de mettre en place des liens retour permettant de revenir à la lecture initiale, à l'image des notes de bas de page mis en place sur Wikipédia par exemple. Toutefois, plusieurs renvois pouvant faire référence à une même section, il aurait été nécessaire de mettre en place plusieurs liens retours afin de ne pas perdre le lecteur. Pour pallier à cela, un système de marque-page numérique a donc été imaginé de telle sorte que lors de l'activation d'un renvoi, une petite icône en marge du texte permet à tout moment de revenir, d'un simple clic, à l'endroit précis où la lecture a été interrompue. Le système ne limite pas le nombre de marque-page que le lecteur peut positionner, il est donc libre de suivre tous les renvois qu'il souhaite sans risque de perdre le fil principal de sa lecture.

Enfin, à l'image des sommaires, table des matières et index, la publication Web de Marmoutier propose un accès permanent au sommaire, à l'index des figures et aux références bibliographiques. Depuis l'index des figures et la liste des références bibliographiques, il est possible d'atteindre toutes les parties du texte où ces éléments sont mobilisés.

3.2. Bonnes pratiques

Depuis l'introduction des principes FAIR en 2016, les pratiques de gestion des données ont changé. Les principes FAIR impliquent que les ressources numériques soient Faciles à trouver, Accessibles, Interopérables et Réutilisables. L'objectif est de faciliter le traitement des données par les systèmes informatiques pour aider les humains à mieux les gérer, leur volume étant en croissance constante. Les bonnes pratiques qui s'appliquent aux données se révèlent aussi valables pour les publications scientifiques.

9. ArSol est le système d'enregistrement de terrain développé par CITERES-LAT depuis 1990 pour l'enregistrement et l'exploitation des données de terrain et des collections archéologiques (http://citeres.univ-tours.fr/spip.php?article505). Il est accessible en ligne depuis 2014 (http://arsol.univ-tours.fr/).

Le respect de ces bonnes pratiques, fortement préconisées par nos tutelles pour progresser vers la Science Ouverte, nous a conduit à choisir d'encoder l'ensemble de la publication au format ouvert de la TEI [10]. Ainsi chaque information de dates, de ressources documentaires, d'éléments archéologiques sont encadrés de balises afin d'être reconnus par la machine. Cela permet en outre de proposer une interaction permanente entre le texte, le plan de localisation des vestiges et la frise chronologique (Fig. 2). Chaque section du texte est associée à une datation qui, durant la lecture, est représentée sur une frise chronologique. En outre, chaque date mentionnée dans le texte est également positionnée sur la frise au survol du pointeur. De même, en fonction de la section affichée à l'écran, le plan de localisation ne mobilise que les vestiges qui sont concernés et, à l'instar des dates, chaque fait archéologique mentionné est localisé spécifiquement sur le plan au survol du pointeur.

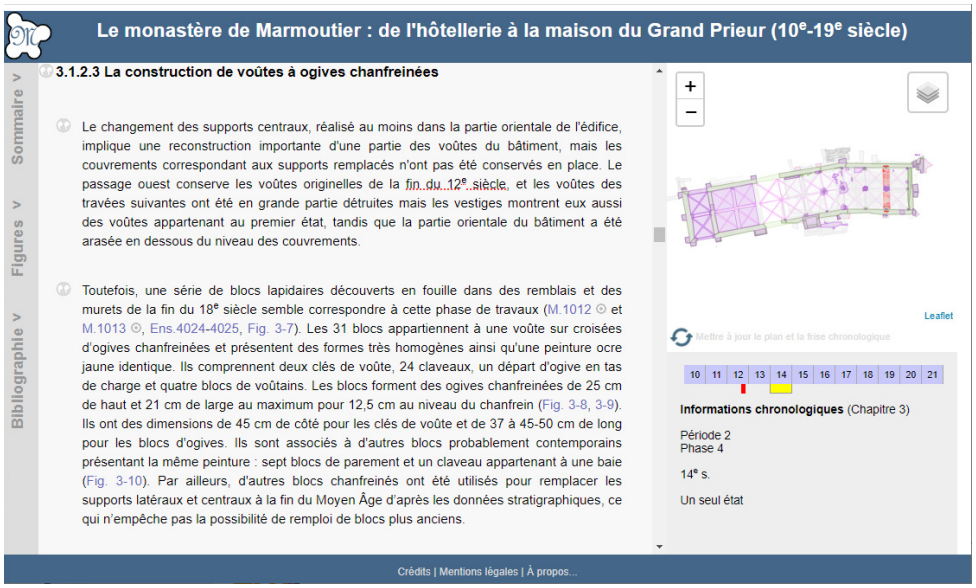


Figure 2: La publication Web des fouilles de l'hôtellerie de Marmoutier (Tours).

Outre l'intérêt de l'exploitation sémantique du contenu du texte encodé en TEI, cela permet également d'accéder à l'information indépendamment de toute application. Le fichier au format TEI est ainsi accessible et lisible tel quel, sans nécessiter d'application spécifique et peut donc être lu par l'homme comme par la machine.

Par ailleurs, à quelque échelle que ce soit, chaque partie du texte (chapitre, partie, sous-partie et paragraphe) est citable. Une petite icône discrète en face de chaque titre ou paragraphe permet d'ouvrir une petite fenêtre indiquant comment le citer et fournissant le lien d'accès direct.

Au-delà de l'emploi de l'ensemble de ces outils pour cette publication, le code est disponible pour son exploitation par d'autres publications scientifiques qui souhaiteraient emprunter la même voie.

10. Text Encoding Initiative : https://tei-c.org/.

Si la publication des fouilles de l'hôtellerie de Marmoutier mobilise des outils numériques pour offrir au lecteur un confort de lecture, elle en reste néanmoins une publication traditionnelle, formant un récit dont la lecture est principalement linéaire, à l'inverse de celle des fouilles du centre paroissial de Rigny qui se distingue par la mise en œuvre du logicisme.

4. La publication logiciste des fouilles du centre paroissial de Rigny

Les fouilles du centre paroissial de Rigny de 1986 à 1989 ont fait l'objet d'une synthèse publiée en 2020 selon les principes du logicisme, élaborés dans les années 1970 par Jean-Claude Gardin [1], et dans un format intégralement numérique.

4.1. Logicisme et publication électronique

Le programme logiciste a été élaboré par Jean-Claude Gardin dans le but de condenser et de schématiser l'architecture des écrits scientifiques. Dès l'origine, il a eu deux objectifs. Le premier était d'ordre épistémologique : il s'agissait de rendre explicites les étapes du raisonnement en distinguant, d'une part, les données de base (ou « propositions initiales »), et d'autre part, les opérations d'inférence effectuées sur ces données pour fonder les hypothèses interprétatives, de manière à constituer une arborescence qui donne une représentation synoptique de l'argumentation et permet d'en évaluer rapidement le bien-fondé [1, p. 244-273]. L'argumentation prend l'allure d'une suite d'opérations d'inférence de P_0 (propositions initiales) à P_n (propositions terminales) en passant par des propositions intermédiaires P_i [9, p. 19] (Fig. 3). Le second objectif était d'ordre éditorial. Comme toute modélisation, la structuration logiciste est une réduction, mais elle conserve la totalité des éléments constitutifs de la construction cognitive, dégagée de l'appareil rhétorique auquel font appel traditionnellement les publications. Elle constitue donc un moyen de réduire le déséquilibre constaté entre volume de production et capacités de consommation bibliographique, et ouvre la voie à une forme de publication adaptée à la prépondérance croissante de la consultation sur la lecture [10].

Dans les décennies 1980-1990, l'analyse logiciste a été expérimentée dans le domaine de l'archéologie, de l'histoire de l'art et de l'histoire, mais sa diffusion est restée très limitée, car l'exercice a longtemps été jugé rébarbatif. Cette première phase à caractère exploratoire a montré l'intérêt épistémologique de l'analyse logiciste mais elle n'a pas eu d'effet concret sur les modes d'édition.

C'est le développement des technologies de l'information qui a permis d'exploiter les possibilités de lecture non-linéaire offertes par les schématisations logicistes, tout en les rendant moins ascétiques grâce à une « mise en scène » multimédia [11]. Le format SCD, acronyme de *Scientific Constructs and Data*, conçu par Valentine Roux (CNRS) et Philippe Blasco (Editions Epistèmes) pour l'édition numérique des réécritures logicistes et des données associées a été utilisé pour des publications en archéologie des techniques : d'abord dans la collection Référentiels (2003-2010), constituée de petits volumes imprimés accompagnés d'un CD-ROM contenant les schématisations logicistes, puis dans la revue en ligne Arkeotek [11], créée par Valentine Roux

11. www.thearkeotekjournal.org.

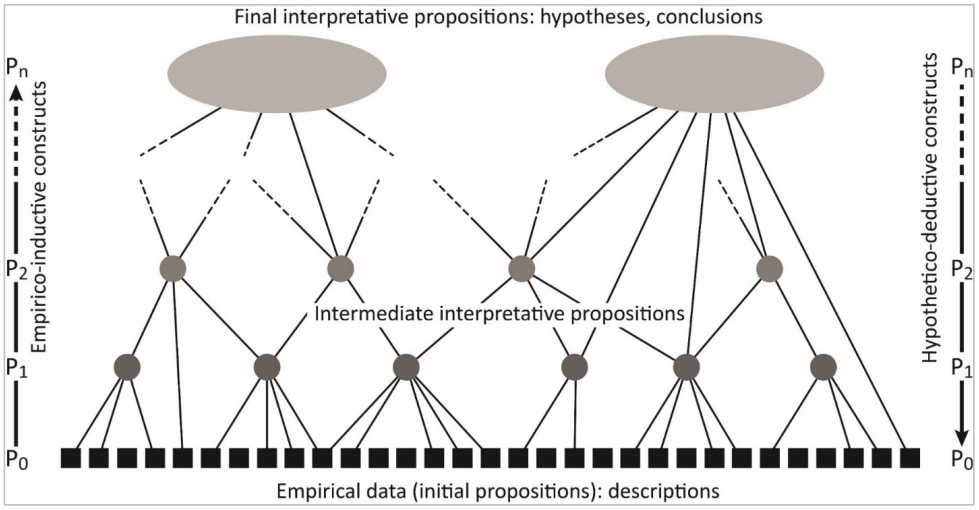


Figure 3: Organisation des propositions logicistes [10, p. 6].

en 2007 pour la publication d'articles et de corpus en archéologie des techniques.

Depuis 2011, grâce au développement des nouvelles technologies du Web, le format SCD a été entièrement reprogrammé en XML-TEI par le Pôle du document numérique de la MRSH de Caen. C'est dans ce format XML-TEI, qui offre des possibilités nouvelles de navigation entre le texte, les schématisations logicistes et les bases de données en ligne, qu'a pu voir le jour en 2020 la publication de la fouille de Rigny, initialement préparée pour une publication dans la collection Référentiels [12].

4.2. L'apport du Web à l'accessibilité des données de fouille

Détruisant son objet d'étude au cours de la fouille, l'acte archéologique est irréversible et ne permet pas de renouveler l'expérience. De ce fait, une fois la fouille terminée, l'enregistrement de terrain constitue la source primaire de l'archéologue et devient particulièrement précieux. Dans les années 1980, le développement de l'informatique a permis le développement et même la multiplication de nombreuses bases de données pour enregistrer les données de terrain. La généralisation d'Internet et l'amélioration des systèmes d'information permet désormais de donner l'accès à ces bases de données de terrain et donc de donner la possibilité à tout chercheur de vérifier une interprétation.

Ainsi, dès 1996, le Royaume-Uni a mis en place l'*Archaeology Data Service* (ADS) [12], centre de ressources numériques de l'université de York donnant accès à l'ensemble des archives de fouilles numérisées (bases de données, documents graphiques, photographies, rapports d'études spécialisées, littérature grise etc.). En outre, la revue *Internet archaeology* [13] a mis en place des publications en ligne dont les liens hypertextes dans le texte renvoient aux données mobilisées,

12. http://archaeologydataservice.ac.uk.
13. http://intarch.ac.uk/.

même si celui-ci reste d'un formalisme classique. En ce qui concerne la numérisation des données de terrain en France, on pourra noter quelques initiatives locales comme la base du Laboratoire Archéologie et territoires de Tours : ArSol [14] (Fig. 4). Depuis 2013, le consortium MASA [15] (Mémoires des Archéologues et des Sites Archéologiques) de l'IR* Huma-Num [16] s'est donné pour objectif d'accompagner les archéologues pour numériser et mettre à disposition leurs archives de fouille en suivant les bonnes pratiques tels que les principes FAIR. À l'échelle européenne, ces publications de bases de données sont en voie de développement grâce au programme ARIADNEplus et la mise en place d'une plate-forme d'accès aux données archéologiques [17]. L'accès en ligne aux données primaires qui permet au lecteur de prendre connaissance de l'enregistrement de terrain constitue une avancée décisive, qui doit permettre d'alléger les publications de fouille des descriptions interminables qu'elles contiennent habituellement et dont seule une faible partie est mobilisée dans l'argumentation.

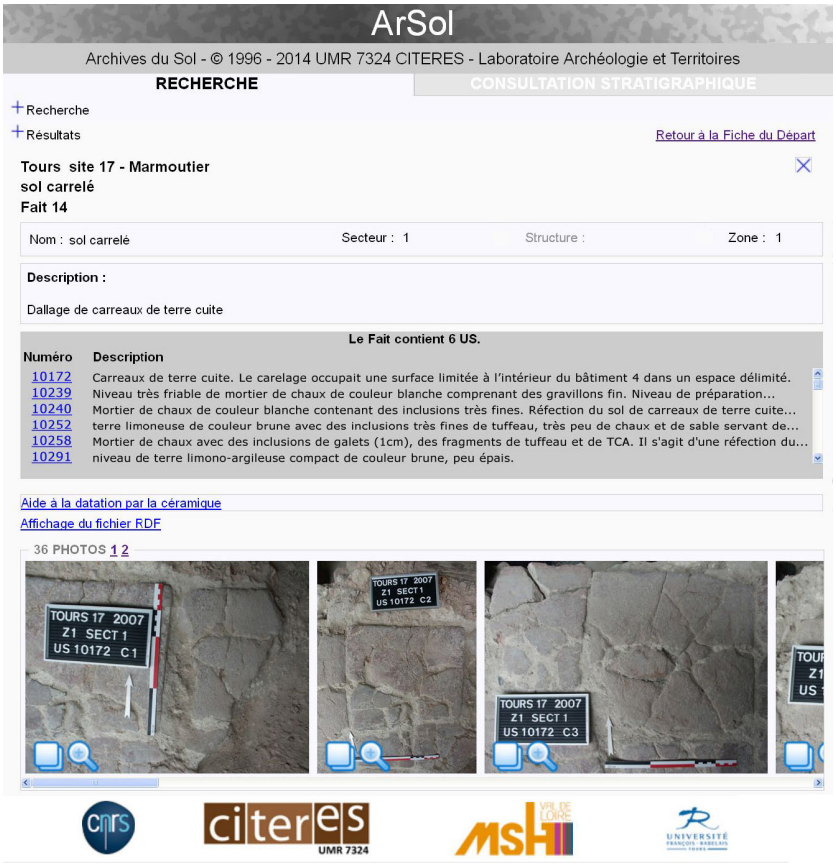


Figure 4: Base de données en ligne ArSol pour les opérations menées par le LAT.

14. http://arsol.univ-tours.fr.
15. https://masa.hypotheses.org.
16. https://www.huma-num.fr/.
17. http://portal.ariadne-infrastructure.eu.

4.3. De la fouille de Rigny à la publication logiciste

L'enregistrement sur le terrain commence par la description des unités stratigraphiques, leur regroupement en entités spatiales hiérarchisées (Faits, Murs Sépultures ou Structures), leur mise en phase, puis leur interprétation fonctionnelle, chronologique et morphologique selon une démarche empirico-inductive. L'écriture logiciste suit l'ordre inverse, partant de l'interprétation pour reconstituer la chaîne des inférences mise en place entre les conclusions P_n et les données de base P_0 mobilisées dans le raisonnement, permettant ainsi de reconstituer à partir des traces observées les modes de vie et les pratiques sociales des hommes (Fig. 5). Les propositions initiales

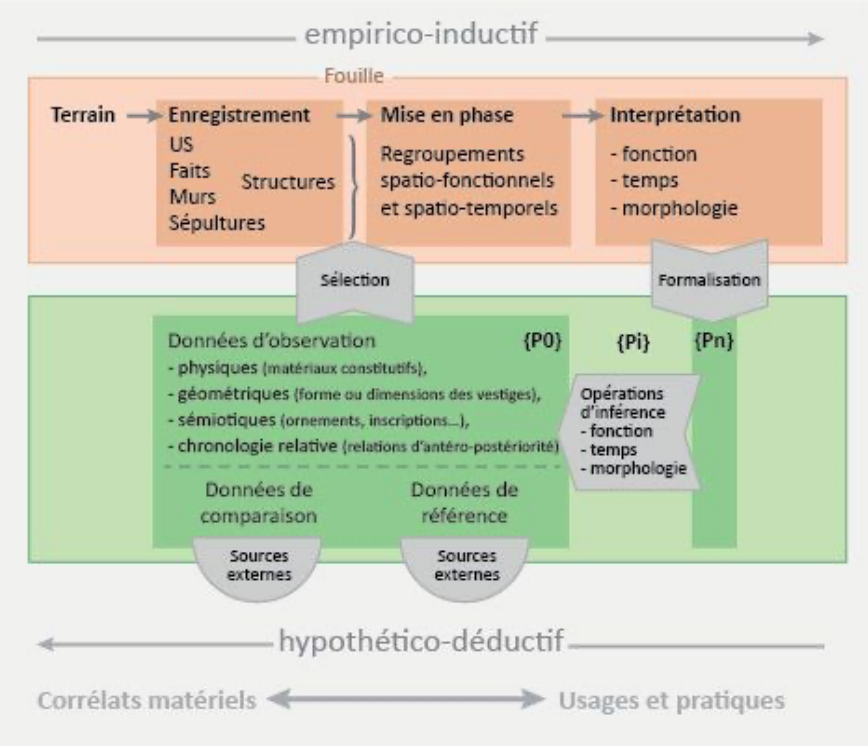


Figure 5: Processus d'interprétation archéologique et de formalisation logiciste.

P_0 se répartissent en trois catégories :

1. Les données d'observation qui sont sélectionnées à partir de l'enregistrement de la fouille. Ces traits descriptifs mobilisés dans l'argumentation peuvent concerner soit des propriétés intrinsèques des entités archéologiques (matériaux, forme...), soit encore leur chronologie relative (relations d'antéro-postériorité).
2. Les données de comparaison ont, comme les données d'observation, le statut de propositions initiales P_0 parce que le constat de ressemblance, qui fonde le raisonnement par analogie, si courant en archéologie, n'est jamais le produit d'une procédure bien définie, mathématique ou logique. Ce constat de ressemblance, une fois déclaré, constitue le fondement de ce que Jean-Claude Gardin appelle le « transfert d'attribut » (« SI deux objets

ou monuments X et Y sont déclarés comparables, au vu de certaines propriétés communes (formes, matériaux, décors etc.), et que Y présente par ailleurs un ou plusieurs attributs connus (date, origine, fonction), ALORS l'on est en droit de transférer à X les mêmes attributs ») [9, 235].

3. Les données dites de référence correspondent à des savoirs tenus pour établis par référence au sens commun [13] ou à des connaissances spécialisées. Entrent dans cette dernière catégorie les analyses de laboratoire (par exemple des datations par le radiocarbone), ainsi que les datations de mobilier lorsqu'elles reposent sur une typo-chronologie établie par d'autres publications. Les données de référence sont considérées comme des données de base P_0, au même titre que les données d'observation et les données de comparaison, parce qu'elles ne font l'objet d'aucune démonstration dans la publication.

Sous leur variété apparente, les opérations d'inférence sont relativement standardisées : à tous les niveaux, depuis les propositions intermédiaires successives (P_1, P_2, ..., P_i) jusqu'aux propositions terminales (P_n), elles consistent presque toujours à attribuer à une ou plusieurs entités une fonction (au sens large du terme), une chronologie (datation, durée...) ou une reconstitution morphologique. Leur compilation dans une base de règles logicistes permettrait de tester leur degré de généralité, et de discuter par exemple des caractéristiques qu'on juge nécessaires pour attribuer à un bâtiment une fonction de stockage, d'habitation ou de lieu de culte dans tel ou tel contexte chrono-culturel.

4.4. Structuration de la publication Web de Rigny

La publication Web des fouilles de Rigny est constituée de plusieurs blocs, qui fournissent autant de niveaux d'accès au contenu, permettant à la fois une lecture rapide et une consultation approfondie. Le premier bloc, intitulé Récit, permet d'avoir accès rapidement, sous forme d'un texte condensé, à l'ensemble des résultats de la fouille, et de consulter si on le souhaite le détail de l'argumentation en accédant, par des liens hypertextes, au bloc qui contient les propositions logicistes, depuis les données (propositions initiales) jusqu'aux propositions terminales (Fig. 6).

Un autre bloc contient les diagrammes logicistes, qui donnent une vision synoptique de l'argumentation sous la forme d'une arborescence qui se développe de gauche à droite. Ces diagrammes, qui permettent une prise de connaissance rapide, sous forme graphique, de l'argumentation, sont interactifs et permettent d'accéder à l'argumentation détaillée (Fig. 7). Celui-ci est produit dynamiquement à partir des textes encodés en XML-TEI et permet d'accéder à toutes les propositions et leurs antécédents et ainsi d'avoir une vue synthétique de l'ensemble des raisonnements. La publication Web offre également la possibilité d'avoir des liens vers la bibliographie, de faire des renvois internes et d'accéder aux fiches d'enregistrement dans la base de données en ligne ArSol. Le lecteur peut ainsi visualiser toutes les chaînes d'inférence dans la structure des diagrammes mais aussi consulter la base de données ArSol contenant les enregistrements de terrain.

Par rapport à une monographie de fouille classique, le recours à l'analyse logiciste entraîne une réduction, sans perte de contenu, du volume du texte, et fait apparaître l'enchaînement des opérations d'inférence en les dégageant de tout appareil rhétorique. La mise en évidence des arguments permet de rendre explicites les articulations du raisonnement et elle facilite la

critique et la comparaison des processus interprétatifs. Elle permet aussi de pratiquer différents niveaux de lecture, depuis la prise de connaissance rapide des résultats jusqu'à l'examen des preuves, et les diagrammes logicistes, qui permettent d'accéder à l'argumentation détaillée, se prêtent bien à une consultation non-linéaire de la publication.

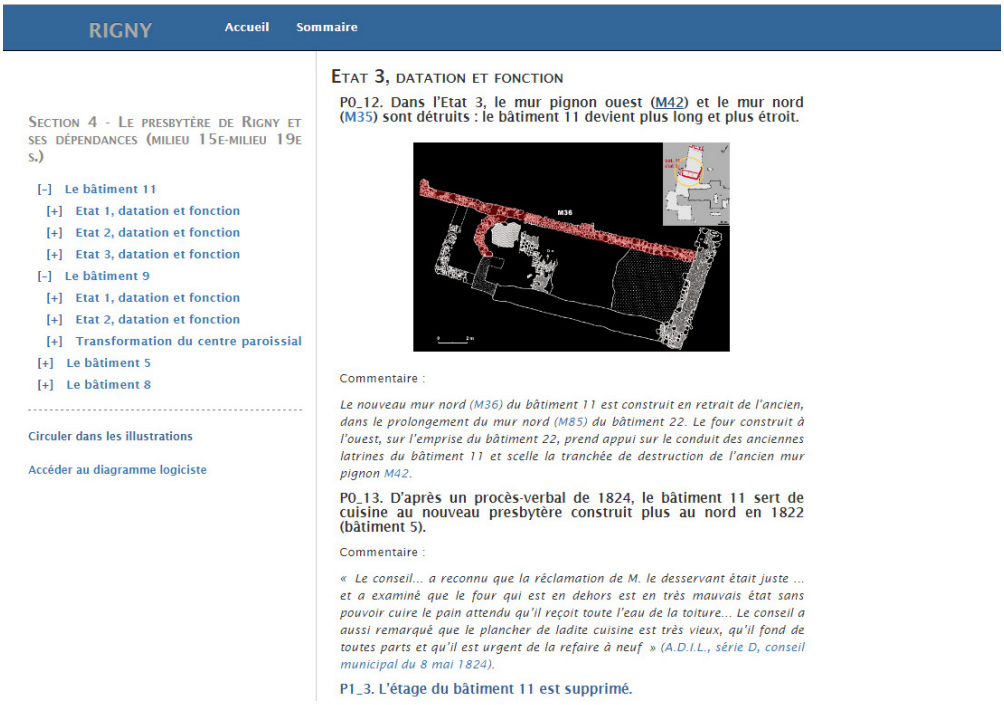


Figure 6: Partie logiciste de la publication des fouilles de Rigny.

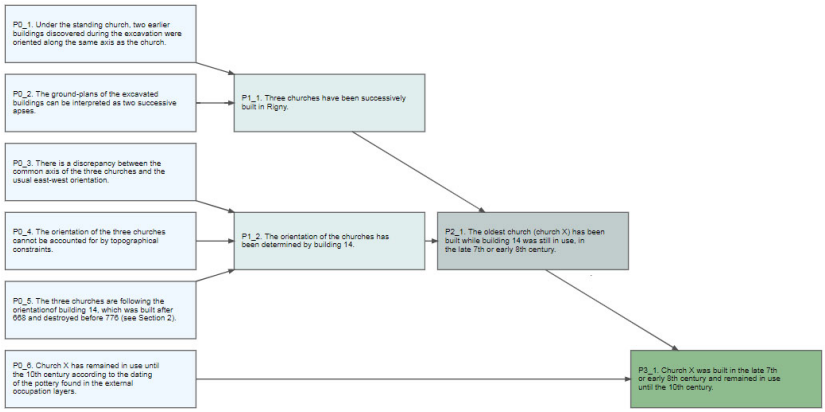


Figure 7: Diagramme logiciste généré automatiquement à partir du texte en XML-TEI.

4.5. Les raisonnements archéologiques dans le Web de données

Puisque toute la publication en elle-même est rigoureusement structurée, son contenu et en particulier les chaînes d'inférences constituent une source de données qui peut être exploitée au même titre qu'une base de données. Ainsi, en procédant à l'appariement des propositions sur les entités du CIDOC CRM [18] et en particulier sur celles du CRMinf [19], nous pouvons assurer l'interopérabilité de cette publication au sein du Web de données [8].

Les propositions initiales P_0 ont été typées selon qu'il s'agit de données d'observation, des données de comparaison, servant de support à un raisonnement par analogie aboutissant à un transfert d'attribut de Temps ou de Fonction, ou des données de référence correspondant à des savoirs tenus pour établis par le sens commun ou par la bibliographie dans le cas de connaissances spécialisées.

Un typage des règles logicistes a également été mise en place en typant les inférences selon qu'il s'agit d'exploiter la Fonction, le Temps ou la Morphologie des éléments faisant l'objet des propositions mobilisées en prémisses. La Fonction est prise au sens large depuis la fonction d'une structure ou d'un bâtiment jusqu'aux interprétations socio-culturelles des usages du site. Le Temps concerne les datations, la chronologie relative, la durée. La Morphologie concerne les hypothèses de reconstitutions architecturales ou les partitions de l'espace (Fig. 8).

Les enregistrements de fouille de la base ArSol sont appariées avec le CRM de base et les extensions CRMSci [20] et CRMArcheo [21] et les chaînes d'inférence sont appariées avec le CRMinf (Fig. 9). Ainsi les données de terrain, rendues interopérables sont mises à dispositions à la fois pour elles-mêmes mais aussi en tant que preuves étayant un raisonnement scientifique.

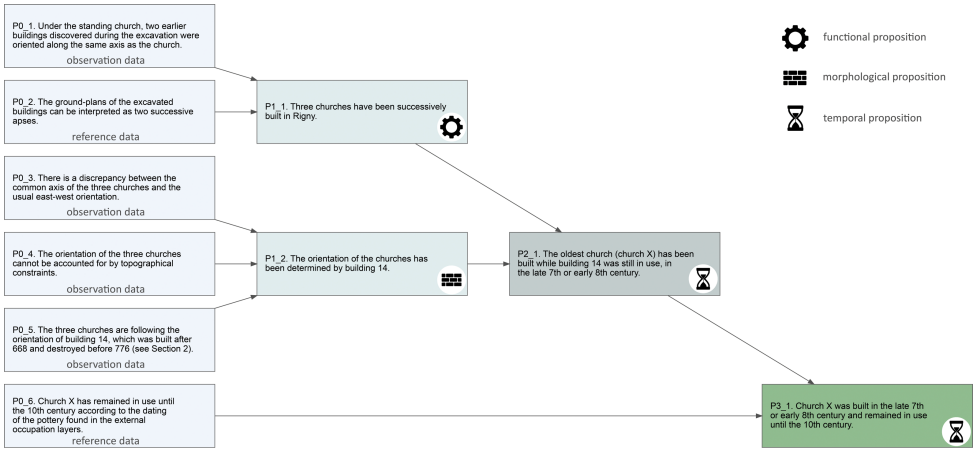


Figure 8: Exemple de typage des propositions.

18. Le CIDOC CRM (Modèle Conceptuel de Référence du Comité International pour la Documentation) est l'ontologie de référence pour modéliser le patrimoine culturel : https://cidoc-crm.org/ (norme ISO 21127 :2014).

19. Le CRMinf est une extension du CIDOC CRM dédié aux inférences et fondé sur la modélisation logiciste : https://cidoc-crm.org/crminf/.

20. Extension du CIDOC CRM pour les observations scientifiques : https://cidoc-crm.org/crmsci/.

21. Extension du CDICO CRM dédié à l'archéologie de terrain : https://cidoc-crm.org/crmarchaeo/.

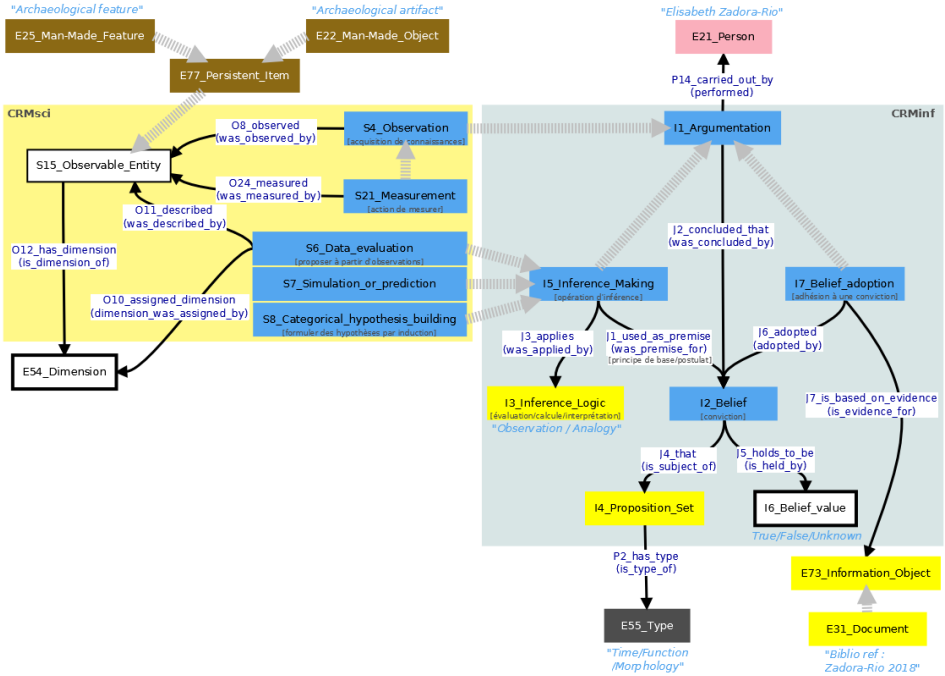


Figure 9: Modèle ontologique explicitant la structuration du CRMinf.

Les propositions initiales P_0 peuvent s'apparier avec les éléments du CRMinf selon qu'elles s'appuient sur des données d'observation et de comparaison ou qu'elles renvoient à des données de références.

L'arborescence des raisonnements et les textes des propositions logicistes sont formalisés dans des fichiers XML-TEI structurés de manière à ce que les balises soient associées explicitement aux entités et propriétés du CIDOC (Fig. 10). Générer un graphe RDF à partir de ces fichiers peut donc se faire aisément avec XSLT.

Ces modèles et appariements sont expérimentaux et nécessitent d'être mis en œuvre avec d'autres sources de données identiques (c'est-à-dire des publications logicistes) pour être éprouvés. Les caractères innovants de la publication de Rigny nécessitent d'être explorés et critiqués avant de pouvoir être déployés pour d'autres fouilles.

Pour le développement de la publication Web de Rigny, le Pôle du Document Numérique de la MRSH de Caen a mis en place un outil permettant de générer automatiquement le graphe des raisonnements en format SVG à partir du fichier source en XML-TEI (Fig. 10). Pour aider les archéologues à exploiter le logicisme pour leurs publications, le consortium MASA a décidé de faire de la rétro-ingénierie et d'exploiter le processus inverse en développant l'application LogicistWriter [22]. Le principe de cette application est de permettre à l'auteur de commencer la rédaction d'une publication logiciste en abordant la construction de l'arborescence de ses raisonnements de manière graphique. Ainsi, l'auteur relie les propositions les unes aux autres depuis les propositions initiales (les preuves) jusqu'aux propositions finales (la synthèse), de

22. https://masa.hypotheses.org/logicistwriter.

```
<div type="chapitre" xml:id="main_div">
        <div type="section1" xml:id="sec1_1">
                <div type="i4_proposition_set" xml:id="section1P0_1" subtype="evidence:observation">
                        <head>proposition title</head>
                        <figure>illustration</figure>
                        <p>comment
                                <ref>< !-- link to arsol data --></ref>
                        </p>
                        <div type="i5_inference_making" subtype="inference:function/time/morphology">
                                <ptr subtype="j1_used_as_premise">premise</ptr>
                        </div>
```

Figure 10: Encodage du fichier TEI en utilisant les entités et propositions du CRMinf.

manière visuelle, sous la forme de graphe. De ce graphe, est généré automatiquement un fichier TEI-XML comportant toute la structure logiciste des raisonnements. Un éditeur XML permet alors de compléter les propositions par du texte, des illustrations, de la bibliographie ou des renvois.

5. Conclusion

Ces publications électroniques ont suivi les développements informatiques en vigueur à un moment donné. Le temps d'appropriation de ces technologies, nécessaire à leur mise en œuvre, n'est jamais négligeable, qu'il s'agisse d'une publication sur CR-Rom ou en ligne, les exigences de l'édition scientifique nécessitant une structuration très différente d'un simple site Web.

À chaque étape, les possibilités offertes par les formats informatiques disponibles ont été exploitées au mieux, afin de proposer des modèles éditoriaux innovants, d'abord en donnant accès à des données complémentaires et structurées sur des annexes numériques (corpus, vidéo), puis en proposant, dans des formats mixtes ou intégralement numériques, une articulation entre texte, argumentaires et preuves. La publication de la fouille du site du château de Tours propose une entrée originale par les données stratigraphiques articulées avec un texte de synthèses et les corpus de preuves mobilisés. Celle des fouilles de l'hôtellerie de Marmoutier offre une interaction dynamique entre tous les éléments de la publication à partir de la lecture linéaire du texte. Enfin, la monographie de fouille du centre paroissial de Rigny met en œuvre l'architecture logiciste élaborée par Jean-Claude Gardin, offrant ainsi des modalités de lecture des résultats et de la démonstration scientifique totalement innovantes.

Chacune de ces expériences constitue une étape qui contribue à renforcer la robustesse éditoriale des productions scientifiques archéologiques, en livrant *in fine* à la fois la synthèse interprétative mais aussi l'intégralité des inférences qui y conduisent, en relation avec les corpus de preuves.

Références

[1] J.-C. Gardin, Une archéologie théorique, Hachette, Paris, 1979.

[2] H. Galinié, P. Husi, J. Motteau (Eds.), Des Thermes de l'Est de *Caesarodunum* au Château de Tours. Le site 3, volume 50 of *Supplément à la RACF*, FERACF, Tours, 2014. Numéro spécial de la collection "Recherches sur Tours", + 1 CD-Rom.

[3] H. Galinié (Ed.), Tours antique et médiéval. Lieux de vie, temps de la ville. 40 ans d'archéologie urbaine, volume 30 of *Supplément à la RACF*, FERACF, Tours, 2007. Numéro spécial de la collection "Recherches sur Tours", + 1 CD-Rom.

[4] P. Husi (Ed.), La céramique médiévale et moderne du Centre-Ouest de la France (11e-17e s.), volume 20 of *Supplément à la Revue Archéologique du Centre de la France*, FERAC, Tours, 2003. + 1 CD-Rom.

[5] P. Husi (Ed.), La céramique du haut Moyen Âge dans le Centre-Ouest de la France : de la chrono-typologie aux aires culturelles, volume 49 of *Supplément à la Revue Archéologie du Centre de la France*, FERACF, Tours, 2013. + 1 CD-Rom.

[6] E. Marot (Ed.), Le monastère de Marmoutier : de l'hôtellerie à la maison du Grand Prieur (10e-19e siècle), volume 77 of *Supplément à la Revue archéologique du Centre de la France*, FERACF, Tours, 2021. URL : http://marmoutier.univ-tours.fr/hotellerie.php.

[7] H. Galinié, P. Husi, X. Rodier, C. Theureau, Élisabeth Zadora-Rio, Arsol, la chaîne de traitement des données de fouilles du laboratoire archéologie et territoires, Les petits cahiers d'Anatole 17 (2005). URL : http://citeres.univ-tours.fr/doc/lat/pecada/F2_17.pdf.

[8] O. Marlet, E. Zadora-Rio, P.-Y. Buard, B. Markhoff, X. Rodier, The archaeological excavation report of rigny : An example of an interoperable logicist publication, Heritage 2 (2019) 761–773. URL : https://www.mdpi.com/2571-9408/2/1/49. doi :10.3390/heritage2010049.

[9] J.-C. Gardin, O. Guillaume, P. Q. Herman, A. Hesnard, M.-S. Lagrange, Systèmes experts et sciences humaines : le cas de l'archéologie, Eyrolles, Paris, 1987.

[10] J.-C. Gardin, La Surproduction des publications en sciences humaines : ses rapports avec la question du mélange des genre, Maison des sciences de l'homme, Paris, 1999. Document de travail publié dans le cadre du séminaire Le Modèle et le Récit, séance du 3 février 1999.

[11] V. Roux, P. Blasc, Faciliter la consultation de textes scientifiques : nouvelles pratiques éditoriales, Hermès et CNRS-Editions, Paris, 2004.

[12] E. Zadora-Rio, H. Galinié (Eds.), L'église de Rigny et ses abords. De la colonia de Saint-Martin de Tours au transfert du centre paroissial (600-1865), Presses universitaire de Caen, Caen, 2020. URL : https://www.unicaen.fr/puc/rigny//accueil.

[13] J.-C. Gardin, The role of local knowledge in archaeological interpretation, in : S. Shennan (Ed.), Archaeological Approaches to Cultural Identity, Unwin Hyman, Londres, 1989, pp. 110–122.

Améliorer la valorisation des données du patrimoine culturel grâce au Linked Open Usable Data (LOUD)

Improving the valorisation of cultural heritage data through Linked Open Usable Data (LOUD)

Julien A. Raemy[1]

[1]*Universität Basel, Digital Humanities Lab, Spalenberg 65, CH-4051 Basel, Suisse*

Abstract

The re-usability, and thus the valorisation, of Linked Open Data (LOD) in the Digital Humanities can be greatly improved by applying the Linked Open Usable Data (LOUD) design principles and the adhering standards which are the Presentation API 3.0 of the International Image Interoperability Framework (IIIF), the W3C Web Annotation Data Model (WADM) and Linked Art, an RDF application profile based on CIDOC-CRM that uses JSON-LD and Getty vocabularies. The various organisations that have taken advantage of these specifications, both technical and social solutions, have greatly contributed to making the most of cultural heritage data and have opened up new possibilities for end users, in particular by meeting the needs of both the scientific community and software developers. As an example, the research project titled "Participatory Knowledge Practices in Analogue and Digital Image Archives" (PIA) is highlighted. This project aims to leverage IIIF, WADM and Linked Art as part of its data model in the context of a Citizen Science initiative focusing on three photographic collections of the Swiss Society for Folklore Studies (SSFS). Thanks to this technological foundation putting LOUD into practice, PIA will enable a more participatory use of the archive, offering possibilities for different target audiences to contextualise, link, contrast and annotate images and their associated metadata.

Keywords

CIDOC-CRM, Application Programming Interface (API), International Image Interoperability Framework (IIIF), JavaScript Object Notation for Linked Data (JSON-LD), Linked Art, Linked Open Data (LOD), Linked Open Usable Data (LOUD), cultural heritage, participatory knowledge practices in analogue and digital image archives, Web Annotation Data Model (WADM)

1. Introduction

L'application du web sémantique et plus précisément du *Linked Open Data* (LOD) ou « données ouvertes liées » dans le domaine du patrimoine culturel se manifeste principalement dans la publication de jeux de données, mises en évidence ici de trois manières, allant de la plus élémentaire à la forme de publication la plus complexe.

Premièrement, les notices descriptives peuvent être accompagnées d'une représentation conforme à la syntaxe *Resource Description Framework* (RDF) à télécharger, par exemple dans des formats de sérialisation RDF/XML ou Turtle. Une deuxième manière est de mettre en place un mécanisme de négociation de contenu permettant autant aux humains qu'aux machines d'accéder aux métadonnées. Enfin, mais de façon plus ponctuelle, lorsque des services SPARQL

. *Workshop on Digital Humanities and Semantic Web*
. julien.raemy@unibas.ch (J. A. Raemy)
. https://github.com/julsraemy (J. A. Raemy)
. 0000-0002-4711-5759 (J. A. Raemy)

sont déployés, ce qui permet de réaliser des requêtes fédérées en ligne ou via un terminal [1].

Ce travail de publication est autant réalisé par les institutions du patrimoine culturel, que sont les bibliothèques, archives et musées, que par les plateformes d'agrégation tel qu'Europeana qui structure les données selon le Europeana Data Model (EDM) lors de leur ingestion [2] voire les agrège directement en RDF lors de projets pilotes [3].

Un autre cas d'utilisation du LOD est celle liée à la curation de données et notamment aux efforts d'alignement de termes descripteurs et d'entités avec des notices d'autorité de bibliothèques telles que Rameau, le Gemeinsame Normdatei (GND) ou encore le Library of Congress Subject Headings (LCSH). Il y a également les services web collaboratifs comme Wikimédia et notamment Wikidata, leur projet de données structurées, qui prend de plus en plus d'importance. Ce dernier est devenu un nœud reliant de nombreux identifiants et une ressource promouvant autant l'utilisation de concepts liés [4] qu'un instrument facilitateur dans les efforts d'alignement, entre autres avec Wikibase, logiciel lui aussi créé sous l'égide de Wikimédia [5].

Les humanités numériques (HN) bénéficient de ces efforts, conduits par les institutions du patrimoine culturel (bibliothèques, archives, musées) et les agrégateurs, principalement pour les deux motifs suivants : améliorer la mise en contexte des données et faire la critique des méthodes et outils appliqués pour les produire. Les HN contribuent même souvent à ces processus, y compris au stade initial, en raison de leur proximité avec ce maillage de communautés. Si les finalités entre les domaines du patrimoine culturel numérique (*Digital Cultural Heritage*) et les HN varient dans une large mesure, ceux-ci partagent des concepts tels que « [...] *mettre à disposition des informations détaillées sur les objets comme fondement de la recherche et un lien étroit avec la création et la perception de la visualisation et de l'imagerie* » [1]. [6].

Hyvönen [7] parle également qu'une transition dans l'utilisation des portails de web sémantique est nécessaire, passant d'une logique de publication de données à celle de l'analyse et de la découverte de connaissances fortuites. Il identifie trois générations de portails :

1. Une première génération de systèmes comprenant le développement de portails pour l'harmonisation des données, l'agrégation, la recherche et la navigation ;
2. Une deuxième génération de systèmes fournissant aux utilisateurs un ensemble d'outils intégrés pour résoudre des problèmes de recherche de manière interactive ;
3. Une troisième génération de systèmes basée sur l'intelligence artificielle qui pourrait résoudre automatiquement des questions de recherche en fonction des contraintes fixées par les scientifiques.

Hyvönen [7] avance que si les HN ont contribué au déploiement de systèmes de la deuxième génération, une demande plus élevée de critique des sources ainsi que des compétences dans l'utilisation d'outils informatiques sophistiqués seront nécessaire pour atteindre la troisième étape afin de stimuler le champ d'action des HN.

Une des clés pour développer de tels systèmes, autant destinés aux humains qu'aux machines et qui puissent être réutilisables facilement, résiderait dans la capacité de faire « partie du web » et non seulement de créer des pièces « sur le web » [8].

1. Traduction de l'auteur

2. Un premier tournant : JSON-LD

Les projets autour du LOD se sont donc principalement concentrés sur la publication et la consommation de données et plutôt dirigés pour un public d'expert·e·s ayant des connaissances en RDF. De plus, les projets LOD ont rarement pris avantage de l'architecture du web, par exemple en construisant des interfaces de programmation applicatives (API - *application programming interface*) ou services web respectant le style d'architecture RESTful.

Un premier tournant a été effectué en 2014 avec la publication de *JavaScript Object Notation for Linked Data* (JSON-LD), une sérialisation RDF qui permet d'exprimer les caractéristiques de RDF de façon à ce qu'elles puissent être interprétées par les développeurs comme du JSON standard, une syntaxe très courante sur le web et qui a pour avantage pour les développeurs de ne pas avoir besoin de comprendre le « formalisme de RDF avant de pouvoir accéder aux données » [9].

Au sein des HN, le développement des API au sein de la communauté IIIF, spécifications toutes sérialisées en JSON-LD, est un exemple de réussite à grande échelle. Les principes de conception promulgués par IIIF ont d'ailleurs servi de base à ceux du *Linked Open Usable Data* (LOUD).

3. Linked Open Usable Data (LOUD)

LOUD ou « données ouvertes liées et utilisables », est un terme et une démarche proposés par Robert Sanderson, impliqué dans la conception des normes et standards présentés ci-après.

Une des premières intentions du LOUD est de permettre autant à la communauté scientifique qu'aux développeurs de logiciels d'accéder aux données. Il convient de trouver un équilibre qui prenne en compte les besoins en matière d'exhaustivité et de précision des données, qui dépend de la construction ontologique, et les préoccupations pragmatiques que sont l'évolutivité et la facilitation d'utilisation.

A l'instar du programme de déploiement en cinq étoiles des données ouvertes liées de Tim-Berners Lee [2], cinq principes de conception (*design principles*) encadrent le LOUD [3] :

A. **La bonne abstraction en fonction du public** (*The right Abstraction for the audience*)

Il faut privilégier les cas d'utilisation au lieu de la rigueur ontologique afin de déterminer le niveau d'interopérabilité.

B. **Peu d'obstacles à l'entrée** (*Few Barriers to entry*)

Les données, et le modèle sous-jacent, doivent être faciles à utiliser et à exploiter. La mise en place de tels systèmes incitera davantage de personnes à y recourir activement.

C. **Compréhensible par introspection** (*Comprehensible by introspection*)

Les données doivent être compréhensibles dans une large mesure simplement en les consultant, sans requérir de l'aide extérieure. Cela peut être achevé en sérialisant les informations en JSON-LD, un format d'encodage de donnée structurées simple à lire et répandu sur le Web.

2. Open Data 5 étoiles : https://5stardata.info/
3. LOUD : https://linked.art/loud/

D. **Documentation comportant des exemples concrets** (*Documentation with working examples*)

Une documentation des plus exhaustives doit être réalisée afin de clarifier la mise en œuvre des cas d'utilisation.

E. **Peu d'exceptions, mais de nombreux modèles cohérents** (*Few Exceptions, instead many consistent patterns*)

Un modèle doit pouvoir contenir le moins d'exceptions possibles pour éviter d'ajouter des règles demandant la création de champs personnalisés au cas par cas.

4. Les standards LOUD

Selon Sanderson [8], il y aurait trois systèmes qui suivent les principes de conception du LOUD : IIIF et plus particulièrement la troisième version de l'API Présentation, le *Web Annotation Data Model* (WADM) du *World Wide Web Consortium* (W3C) et *Linked Art*. Ces trois spécifications sont complémentaires et peuvent être utilisés séparément ou conjointement.

Si l'on considère la relation entre les principes FAIR (*Findable, Accessible, Interoperable, Reusable*) [10] et LOUD, il apparaît que les premiers sont liés à l'environnement dans lequel se trouvent les données et que le second concerne le contenu en soi. En décomposant l'acronyme LOUD, les termes *Linked*, *Usable* et *(machine-readable) Data* peuvent être considérés comme des caractéristiques des données (et de leur utilisation une fois transférées dans un environnement). *Open* peut constituer en quelque sorte le pendant du principe de réutilisabilité énoncé dans FAIR [11].

4.1. International Image Interoperability Framework (IIIF)

Le « cadre international d'interopérabilité des images » (IIIF) est une initiative communautaire, réunissant d'importants acteurs des domaines académique et culturel, qui a défini plusieurs API partagées permettant de normaliser la manière dont les ressources iconographiques et audiovisuelles sont décrites et diffusées sur le web. L'implémentation des API de IIIF permet aux institutions de mieux valoriser leurs collections numérisées ou nées numériques en offrant, par exemple, des possibilités de zoom profond, de comparaison, de recherche plein texte d'objets océrisés ou encore d'annotation [12]. Depuis mars 2022, il existe désormais six API stables qui sont conçues et approuvées par le consortium et la communauté IIIF [4]. Les deux spécifications principales sont les API Image et API Présentation. La première étant un « *service web pour manipuler une image à travers une URL*» et la seconde « *spécifie les informations nécessaires à la présentation d'un objet numérique* » [13].

IIIF a établi des principes de conception [5] pour guider la manière dont les spécifications sont créés, en veillant notamment à ce que le périmètre de ces travaux soit constitué autour de cas d'utilisation communs, en évitant de dépendre de technologies qui constituent une barrière à l'entrée et en étant conforme aux normes d'architecture web du W3C et de l'*Internet Engineering*

4. API Specifications - International Image Interoperability Framework : https://iiif.io/api/
5. IIIF Design Principles : https://iiif.io/api/annex/notes/design_principles/

Task Force (IETF). Par exemple, lors des sorties des troisièmes versions des deux API principales en 2020, IIIF a suivi les évolutions du W3C et de JSON-LD en intégrant JSON-LD 1.1, sortie la même année, et de s'appuyer sur le WADM au lieu de *Open Annotation*.

Le modèle conceptuel de IIIF et de l'API Présentation se base sur le *Shared Canvas Data Model* décrivant la représentation numérique d'un objet physique au sein d'un canevas auquel est attribué des dimensions pouvant accueillir le contenu, autrement dit une ou plusieurs images ou annotations associées [14]. Ce modèle a été réalisé lors des réunions du *Digital Manuscript Technical Working Group*, groupe actif entre 2010 et 2013, qui fut une initiative préliminaire à IIIF réunissant principalement des médiévistes à l'initiative de Université de Stanford [13] [6].

Qui plus est, si IIIF ne se base pas *stricto sensu* sur RDF, il respecte sa syntaxe dans un sens conceptuel puisqu'il est en quelque sorte « *un support visuel pour le LOD* » [7] [17]. Si des assertions RDF peuvent bien être créés à partir de ressources IIIF, ce n'est en tout cas pas un but recherché de la communauté. IIIF permet de faire pointer vers d'autres données structurées ou représentations différentes d'un objet via la propriété `seeAlso`, mais il n'existe pas encore de consensus ou de bonnes pratiques sur la manière dont cela devrait être effectué à des buts d'agrégation même si cela pourrait changer avec, premièrement l'API Change Discovery sortie en 2021 et la création du groupe communautaire Discovery for Humans [18, 19]. Néanmoins, des expérimentations ont déjà été réalisés, notamment par Europeana [3, 20].

4.2. Web Annotation Data Model

Le WADM a été créé parallèlement à un vocabulaire et un protocole par un groupe dédié du W3C qui a retravaillé la spécification Open Annotation, développée en 2013, « *(...) définissant un cadre interopérable pour la création d'associations entre des ressources connexes, des annotations, à l'aide d'une méthodologie conforme à l'architecture du Web* » [8] [21].

Le WADM est, comme tous les standards LOUD, sérialisé en JSON-LD. Son principe de fonctionnement, comme illustré sur la Figure 1, repose sur la division d'une annotation en deux parties distinctes que sont le corps (`body`), qui correspond à la ressource sur laquelle on cherche à annoter quelque chose, et la cible (`target`), qui représente l'objet que l'on annote.

Le contexte de création des annotations est défini par le biais d'une dizaine de « motivations » qui seront interprétées selon le client. Par exemple, le visualiseur d'images Mirador 3, qui est compatible aux API de IIIF et au WADM, utilisera et affichera par défaut les annotations ayant *commenting* comme motivation, un de cas d'utilisation les plus fréquents.

4.3. Linked Art

Linked Art est un profil RDF de l'ontologie CIDOC-CRM sérialisé en JSON-LD qui se veut pragmatique et moins complexe que CIDOC-CRM en mettant à disposition une API et qui

6. La plupart des ressources compatibles à IIIF restent à ce jour des manuscrits et livres anciens. Premièrement, car il s'agit en quelque sorte du cœur de la communauté IIIF qui a réellement débuté en 2014 lors de la publication de la deuxième version de e-codices, bibliothèque virtuelle des manuscrits en Suisse, avec le déploiement des deux API principales [15]. Dans un deuxième temps, énormément d'argent et d'effort ont été investi par les institutions dans la numérisation de ces objets, tombés dans le domaine public depuis longtemps et qu'il est plus aisé de disséminer [16].
7. Traduction de l'auteur
8. Traduction de l'auteur

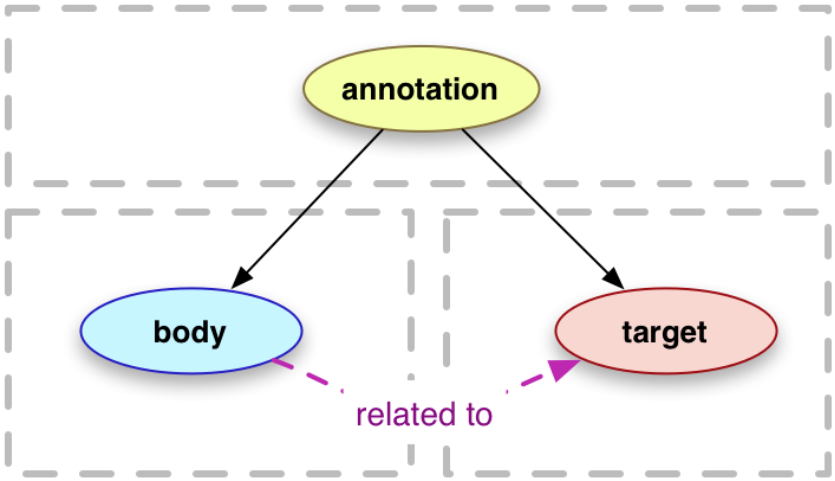


Figure 1 : Principes de fonctionnement du WADM

fait appel aux vocabulaires du Getty (tels que AAT et ULAN) pour préciser les entités des ressources du patrimoine culturel décrites. Linked Art utilise également d'autres ontologies RDF couramment utilisées comme RDFS ou Dublin Core pour compléter son modèle, par exemple lorsqu'il faut désambiguïser des noms de propriétés proches ou similaires déjà utilisés par CIDOC-CRM.

Linked Art décrit les assertions dans un paradigme axé sur les événements, plutôt que sur les objets (comme c'est le cas avec la plupart des normes de bibliothèques ou encore avec Dublin Core ou Schema.org) et est documenté, tout comme IIIF, dans une démarche ascendante où les cas d'utilisation communs des parties prenantes influencent grandement le modèle. Techniquement, voici comment Linked Art fonctionne [9] :

« *Linked Art utilise le concept* `contexte` *de JSON-LD pour faire correspondre les classes sélectionnées et leurs relations respectives à partir de l'ontologie RDF publiée de CIDOC-CRM vers des propriétés JSON en appliquant un ensemble de règles établies.* » [9].

Parmi ces règles, ou patterns, les suivantes méritent d'être mentionnées :

— **Spécificité des classes et classifications** (*Types and Classifications*) : CIDOC-CRM est un environnement qui doit être enrichi de vocabulaires et d'ontologies additionnels pour être fonctionnel. Le mécanisme fourni pour ce faire est la propriété `classified_as`, qui fait référence à un terme d'un vocabulaire contrôlé. Ceci est en contraste avec la propriété `type`, qui est utilisée pour les classes définies par CIDOC-CRM (`crm:P2_has_type`). Ainsi, la responsabilité du maintien de la structure des classes est déplacée de l'ontologie vers le vocabulaire [9].

— **Noms et identifiants d'une ressource** (*Names and Identifiers for a Resource*) : La désignation des objets se fait via la propriété `_label` (`rdfs:label`) pour la documentation interne. En revanche, pour leur donner un nom ou un label si aucun identifiant n'a été créé via un système, il faut utiliser la propriété `_identified_by` en indiquant que la ressource

9. Traduction de l'auteur

a le type `Name`. Ce pattern a pour objectif de rendre l'instance du nom comme unique.

— **Partitionnement d'entités** (*Parts*) : Linked Art a choisi le partitionnement d'entités via la propriété `part` pour tous les cas où l'on souhaite regrouper des propriétés spécifiques d'une entité. Ceci dans le but d'obtenir des descriptions de plus en plus granulaires ou spécifiques. L'appartenance à un set est traité différemment dans le cas où cet ensemble pourrait ne plus comporter de membres et toujours en être un (on peut imaginer une institution où tous les employés décident de se retirer). Dans ce cas là, les propriétés `member` ou `member_of` sont utilisés.

Par ailleurs, Un avantage non négligeable du modèle est la possibilité d'intégrer des objets et services numériques [10], de sorte qu'il est aisé de pointer vers des ressources IIIF en donnant des informations quant à leur niveau de conformité par rapport à une API au sein d'un modèle de données [22].

Linked Art est une initiative encore récente et en développement. Néanmoins, le modèle a notamment été adopté par Pharos, le consortium international des archives photographiques, pour mettre à disposition ses descriptions de collections [23] et est en train d'être implémenté par la Galerie nationale d'art des États-Unis (*National Gallery of Art*) ou encore l'université de Yale pour uniformiser leurs pratiques en matière de publication de données.

5. Pratiques de connaissance participatives dans les archives d'images analogiques et numériques (PIA)

Dans ce chapitre sont présentés le projet de recherche PIA, la mise en œuvre globale des principes de conception et des standards LOUD suivis de trois cas d'utilisation décrits de manière plus détaillée, ainsi que les perspectives d'avenir et l'impact que les standards LOUD peuvent apporter lors de leur déploiement dans le cadre d'un projet de science citoyenne comme celui-ci.

5.1. Bref descriptif du projet PIA

En février 2021, le séminaire d'anthropologie culturelle et d'ethnologie européenne (*Seminar für Kulturwissenschaft und Europäische Ethnologie*) et le laboratoire d'humanités numériques (DHLab) de l'Université de Bâle ainsi que le *Institute of Design Research* de la Haute école des arts de Berne ont démarré un projet de recherche interdisciplinaire intitulé « Pratiques de connaissance participatives dans les archives d'images analogiques et numériques » (PIA - ***Participatory*** *Knowledge Practices in Analogue and Digital* ***Image Archives***) [11].

Le projet PIA, qui se déroule de février 2021 à janvier 2025, prévoit la mise en place de processus, d'outils et d'interfaces permettant de générer et rendre visible la connaissance de manière participative en permettant notamment un accès et une exploration intuitifs, basée sur l'exemple de trois collections des archives photographiques [12] de la Société suisse des traditions populaires (SSTP) - *Schweizerische Gesellschaft für Volkskunde* (SGV) - qui rassemblent des

10. Linked Art - Digital Integration : https://linked.art/model/digital/
11. PIA Project : https://about.participatory-archives.ch/
12. Fotoarchiv der Schweizerischen Gesellschaft für Volkskunde : https://archiv.sgv-sstp.ch/

photographies sur des sujets divers tels que la vie quotidienne, la tradition et l'identité, les formes de travail et d'habitation [22].

Parallèlement au déploiement de l'infrastructure et au design de ces interfaces, le projet étudie les phases des archives analogiques et numériques dans une perspective d'anthropologie de la connaissance, de technique et de communication.

Les nouvelles collectes de données font également partie d'une utilisation participative des archives : les utilisatrices et utilisateurs mettent à disposition leurs propres images, ce qui permet d'élargir les collections, elles et ils écrivent ou racontent des faits intéressants à propos des objets ou font part de leurs propres impressions, expériences et analyses. Toutes ces données peuvent à leur tour faire partie des archives et contribuer ainsi, par leur recontextualisation, à la production de connaissances dans différents domaines. Ainsi, les thèmes d'une collection peuvent également être actualisés et étudiés sous de nouveaux angles.

L'objectif du projet est d'analyser et de décrire systématiquement les pratiques archivistiques historiques et actuelles ; de faire des archives un lieu vivant de création, préservation et transmission des connaissances. PIA veut s'inscrire dans la mouvance de l'ouverture des données et afin d'y parvenir, sept priorités ont été identifiés au sein d'une vision commune prenant en compte les aspects interdisciplinaires du projet :

1. Accessibilité ;
2. Hétérogénéité ;
3. Matérialité ;
4. Interopérabilité ;
5. Interconnexions ;
6. Intelligence artificielle ;
7. Gestion des biais.

5.2. Application des principes de conception et des standards LOUD

Si instinctivement, les principes de conception encadrant le LOUD s'alignent et peuvent plus facilement contribuer à la réalisation des priorités 1, 4 et 5, toutes les priorités du projet PIA peuvent en fait être réalisés ou optimisés par le biais de Linked Art, IIIF ou le WADM.

Plus concrètement, PIA utilisera les trois standards LOUD mentionnés ci-avant. Dans un premier temps, les deux API principales de IIIF vont être déployées, ce qui permettra, entre autres, aux utilisatrices et utilisateurs d'annoter les ressources conformément au WADM. Vu qu'il s'agit d'une initiative promouvant la science citoyenne ou participative [13], il a par exemple été convenu d'autoriser le téléversement d'images privés sur l'interface utilisateur pour contextualiser, comparer, complémenter ou encore déconstruire le corpus de la SSTP, obtenu via l'API DaSCH Service Platform (DSP), service du DaSCH qui est actif dans la préservation à long-terme des données de sciences humaines en Suisse.

Les métadonnées, en plus d'une sérialisation en Schema.org, seront également exposées via le modèle Linked Art [25]. Les différentes informations seront reliées dans l'API centrale de PIA, par le biais de la propriété `seeAlso` au sein des ressources IIIF pointant vers une représentation de l'objet en Linked Art, en ayant des listes d'annotations WADM rattachées aux ressources IIIF et enfin en indiquant au sein du modèle Linked Art lorsqu'un service web IIIF est disponible.

13. Au sein du titre de travail de la future thèse de doctorat de l'auteur, le terme *Citizen Humanities* - « science citoyenne dans les sciences humaines » - est employé pour le démarquer des initiatives de science citoyenne émanant des sciences naturelles [24]

Cet « écosystème LOUD », schématisé dans la Figure 2, permettra de répondre aux besoins de différents publics, que ce soit des humains ou des machines.

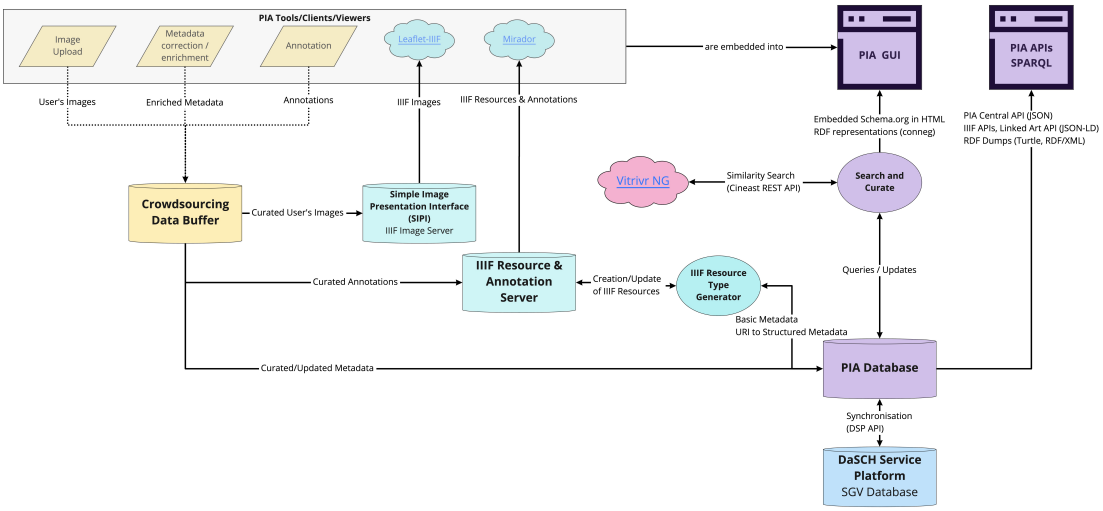


Figure 2 : Panorama simplifié de l'architecture de PIA

Il convient donc de déployer des efforts particuliers dans la modélisation des différents cas d'utilisation, mis en évidence dans les sous-sections successives par le biais d'une photographie issue de la collection Ernst Brunner (cf. Figure 3), et de pouvoir les formaliser tout en gérant au mieux les flux entre les différents services et logiciels de l'infrastructure.

Figure 3 : SGV_12N_02727 (Alte Bildnummer : BC 27). [Menschenmasse und Propellerflugzeuge]. Dübendorf, undatiert. Negativ schwarz/weiss 6x6cm. SGV_12 (Ernst Brunner). Schweizerische Gesellschaft für Volkskunde (SGV). https://archiv.sgv-sstp.ch/resource/424962.

5.2.1. Cas d'utilisation 1 : modélisation d'une photographie numérisée et gestion de ses identifiants

La modélisation en Linked Art des différentes entités (objet, personne, événement, etc.) se basera sur la sélection de ressources qui puissent représenter un ensemble cohérent. Pour la ressource qui nous intéresse, une numérisation à partir d'un négatif noir et blanc, celle-ci est très représentative de la collection Ernst Brunner comportant plus de 47'000 objets ayant cette typologie.

Tout d'abord, il a fallu choisir comment un tel objet allait être représenté au centre de la modélisation, de manière conceptuelle, physique ou numérique. Autrement dit est-ce qu'il s'agit plutôt de la photographie telle qu'elle nous est présentée sur le site de la SSTP, du négatif conservé au sein des archives ou encore du fichier issu de la numérisation, réalisée il y a quelques années par le DHLab, à partir du négatif. C'est la dernière option qui a été retenue, car la première était difficilement envisageable vu qu'il s'agit plutôt d'une reconstruction intellectuelle de la photographie et qui serait plus complexe à modéliser en Linked Art et la deuxième nous paraissait superflue à sérialiser car l'utilisation majeure de cette ressource est désormais numérique.

```
{
  "@context": "https://linked.art/ns/v1/linked-art.json",
  "id": "https://linkedart.participatory-archives.ch/object/14759",
  "type": "DigitalObject",
  "_label": "PIA ID 14759 - [Menschenmasse und Propellerflugzeuge]",
  "classified_as": [
    {
      "id": "http://vocab.getty.edu/aat/300215302",
      "type": "Type",
      "_label": "Digital Image"
    }
  ],
```

Extrait 1 – Modélisation Linked Art de la photographie comme `DigitalObject`

En d'autres termes, plutôt que d'utiliser la classe CIDOC-CRM `HumanMadeObject`, la modélisation débute avec une instanciation de la classe CRMdig `DigitalObject` en suivant notamment les consignes promulguées pour la modélisation de l'entité Object le site web de Linked Art, par exemple en mettant en avant le vocabulaire AAT (cf. Extrait 1).

S'il a été décidé de modéliser à partir de la photographie numérisée, son contexte de création est évidemment évoqué (cf. Extrait 2). Au sein du fichier JSON-LD complet - encore en phase de développement - des métadonnées concernant le lieu où la photographie a été prise ainsi que sur ses dimensions physiques viennent enrichir la description du négatif.

```
  "created_by": {
    "type": "Creation",
    "_label": "Digitisation of Photograph",
    "used_specific_object": [
      {
        "type": "HumanMadeObject",
        "_label": "Negative of [Menschenmasse und Propellerflugzeuge]",
        "classified_as": [
```

```
          {
            "id": "http://vocab.getty.edu/aat/300128343",
            "type": "Type",
            "_label": "Black and White Negative",
            "classified_as": [
              {
                "id": "http://vocab.getty.edu/aat/300435443",
                "type": "Type",
                "_label": "Type of Work"
              }
            ]
          }
        ],
        ...
      }
  ]
},
```

Extrait 2 – Contexte de création et lien vers le format original

Cette ressource comporte plusieurs identifiants :

— `BC 27` : identifiant attribué par le photographe Ernst Brunner
— `SGV_12N_02727` : identifiant attribué par les archives de la SSTP
— `14759` : identifiant créé pour les besoins du projet et retenu comme point d'entrée principal.

En plus de ceux-ci et pour suivre l'un des patterns de Linked Art, s'ajoute l'identification par le nom, ici du titre qui a été attribué par la SSTP. Vu que chaque identifiant a été créé de manière distincte (logique propre, *numerus currens* de manière manuelle et identifiant assigné automatiquement par un système informatique), chaque nœud pointe vers un terme AAT différent et a une labélisation propre (cf. Extrait 3).

En outre, Il faut encore noter que cette ressource a aussi un numéro dédié sur le site web des archives de la SSTP (`424962`) et que prochainement, elle obtiendra un nouvel identifiant lors de la migration de la base de données qui devrait être effectué par le DaSCH dans le courant de l'année 2022. Néanmoins, ces deux hyperliens sont ou seront plutôt traités comme des représentations complémentaires, le premier est d'ailleurs aussi modélisé dans le fichier JSON complet via la propriété `subject_of`.

```
"identified_by": [
  {
    "type": "Name",
    "content": "[Menschenmasse und Propellerflugzeuge]",
    "classified_as": [
      {
        "id": "http://vocab.getty.edu/aat/300404670",
        "type": "Type",
        "_label": "Owner-Assigned Title"
      }
    ],
    "language": [
      {
        "id": "http://vocab.getty.edu/aat/300388344",
        "type": "Language",
        "_label": "German"
      }
    ]
  },
  {
    "type": "Identifier",
    "content": "BC 27",
    "classified_as": [
      {
        "id": "http://vocab.getty.edu/aat/300417447",
        "type": "Type",
        "_label": "Creator-Assigned Number"
      }
    ]
  },
  {
    "type": "Identifier",
    "content": "SGV_12N_02727",
    "classified_as": [
      {
        "id": "http://vocab.getty.edu/aat/300312355",
        "type": "Type",
        "_label": "SGV Signature"
      }
    ]
  },
  {
    "type": "Identifier",
    "content": "14759",
    "classified_as": [
      {
        "id": "http://vocab.getty.edu/aat/300404621",
        "type": "Type",
        "_label": "PIA ID"
      }
    ]
  }
],
```

Extrait 3 – Nom et identification de l'objet `14759`

5.2.2. Cas d'utilisation 2 : dissémination des ressources numériques et intégration au modèle de données

Afin de valoriser les ressources numériques et de les disséminer de manière standardisée, les deux API principales de IIIF ont été déployées au sein de l'infrastructure PIA.

Premièrement, chaque image est servie via notre instance de serveur Simple Image Presentation Interface (SIPI), développé et maintenu par le DHLab. Autant la collection photographique de la SSTP que les futurs fichiers téléversés par les utilisatrices et utilisateurs sont et seront converties dans le format JPEG2000 pour générer les différentes tuiles permettant un zoom profond. En parallèle, SIPI s'assure de générer également les informations nécessaires (*Image Information Request URI Syntax*) pour une compatibilité à l'API Image 3.0 [26].

Ensuite, sur la base d'un modèle de `Manifest`, ressource IIIF principale au sein de l'API Présentation [27], nous avons généré à l'aide d'un script, pour tous les objets disponibles dans notre base de données, des fichiers JSON-LD en liant les données ainsi que les métadonnées descriptives et légales. Pour notre ressource, il s'agit de https://iiif.participatory-archives.ch/14759/manifest.json - fichier qui va continuer d'être enrichi, notamment en pointant vers plus de métadonnées structurées [14].

Vu que notre API Linked Art est en phase de développement, il n'y a pas encore d'intégration au sein des Manifestes IIIF, par le biais de la propriété `seeAlso`, pointant vers une description Linked Art des objets. Dans l'autre direction, la Figure 4 illustre une intégration des deux API de IIIF au sein du modèle de données en suivant les règles de bonnes pratiques édictées par la communauté Linked Art, par exemple en indiquant la version des API, leur format (JSON-LD) ainsi qu'en allouant une classe à chaque nœud [15].

5.2.3. Cas d'utilisation 3 : enrichissement des métadonnées et annotation des contenus

Un des buts du projet PIA est de permettre l'enrichissement et la correction des métadonnées descriptives par le biais d'un approvisionnement par la foule (*crowdsourcing*). Ceci pourrait être réalisé par l'intégration ou l'utilisation d'outils *Open Source* conformes à IIIF tels que Madoc ou Zooniverse et avec lesquels il est possible de modifier les métadonnées en important des ressources IIIF ou de les annoter directement [28, 29].

Au sein de la ressource `14759`, la date est par exemple inconnu même s'il s'agit assez certainement d'une photographie prise à la fin des années 1930 en suivant la logique de numérotation du photographe. De même, sur l'un des avions, on peut apercevoir une inscription portant le nom « József Kiss ». Il reste par contre à savoir s'il s'agit du héros militaire hongrois de la première

14. A noter que nous nous efforçons de faire en sorte que les liens des ressources IIIF restent stables et sommes en train de développer une politique d'attribution d'identifiants pérennes à l'aide du schème *Archival Resource Key* (ARK). Cf. https://github.com/Participatory-Image-Archives/ark

15. Un graphe complet au format SVG a été généré en parallèle de la sérialisation en JSON-LD.

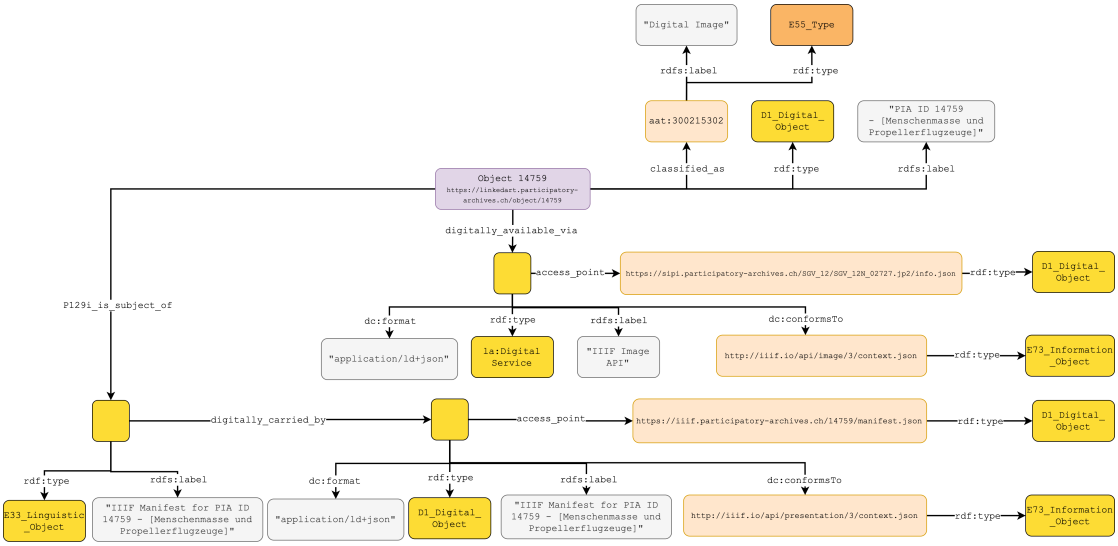


Figure 4 : Intégration de ressources/services IIIF au sein d'une modélisation Linked Art

guerre mondiale (https://www.wikidata.org/entity/Q385425) ou de l'un de ses homonymes, notamment le poète de la fin du XIXe siècle (https://www.wikidata.org/entity/Q920515).

Grâce à l'outil Annonatate, permettant de créer des annotations conformes au WADM sur la base de ressources IIIF [16], il a été assez facile de sélectionner une partie d'image et de l'annoter (cf. Figure 5). Cette annotation a généré le fichier JSON-LD suivant : https://julsraemy.github.io/annonatate/annotations/14759-p1-list.json.

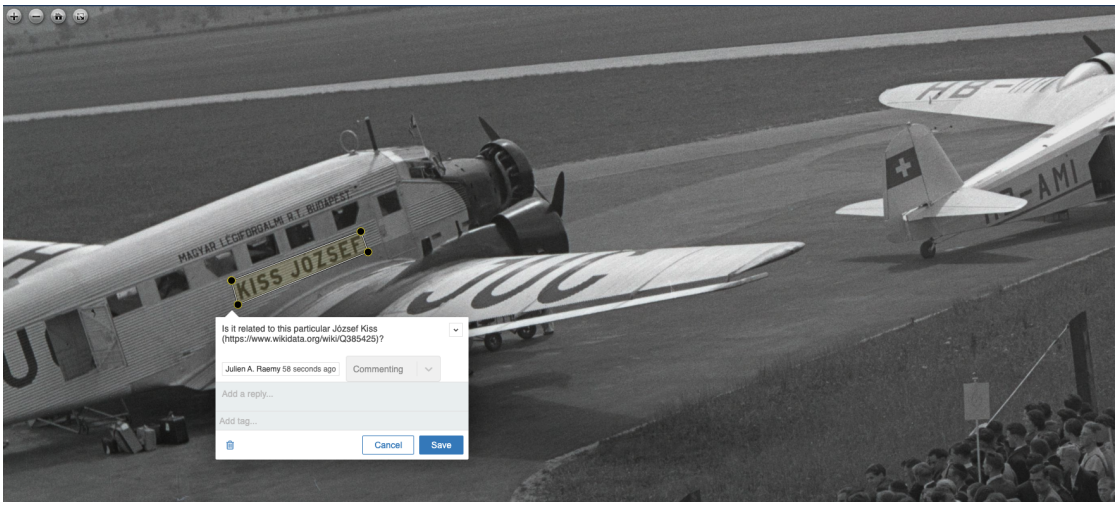


Figure 5 : Annotation du Manifeste IIIF `14759` au sein d'Annonatate

16. Annonatate permet une sélection fine des `motivations` ainsi que l'ajout de mots-clés associés. Il est donc théoriquement possible de recourir à un vocabulaire contrôlé.

Ce processus d'annotation a surtout été une démonstration de faisabilité et à quel point ces standards permettent de plus facilement échanger des données sur le Web. Pour le projet de recherche PIA, il s'agira, entre autres, de trouver des solutions techniques et organisationnelles pour modérer les flux d'information.

5.3. Perspectives

Pour chacun des cas d'utilisation présentés dans cet article, l'un des trois standards LOUD a été davantage sollicité (dans l'ordre : Linked Art, IIIF, puis le WADM). Par rapport à une implémentation complète du modèle de données et des composantes LOUD, en plus de la modération des flux mentionnés ci-avant, il y a encore des incertitudes sur la création de modèles et de règles sous-jacentes pouvant accommoder les différentes typologies d'entités. Il faudra donc repenser les processus au sein de notre prototype et voir comment modéliser les cas d'utilisation de la manière la plus pragmatique possible, ce qui est déjà le cas avec la modélisation de la ressource `14759` en Linked Art où des choix ont été opérés pour garder la numérisation et le négatif au sein d'un même fichier pour les garder connectés sans être tout à fait correct d'un point de vue ontologique.

Ce qui est déjà certain c'est que cet écosystème LOUD est déjà un véritable actant - à prendre au sens latourien du terme - dynamique de l'infrastructure et influence grandement la façon dont les informations peuvent être sauvegardées et disséminées mais également à quel point cela va impacter les utilisatrices et utilisateurs des interfaces [30].

Dernièrement, il restera bien évidemment à étudier l'évolutivité d'un tel système et voir à quel point ces standards ouverts et maintenus par des communautés ou bâties autour d'une pratique communautaire peut servir d'exemple dans les HN et ainsi devenir ce que Hyvönen [7] mentionne comme cette troisième génération de portails pouvant résoudre automatiquement des questions de recherche en fonction des contraintes fixées par les scientifiques tout en gardant à l'esprit les besoins et apports d'un plus large public.

Conclusion

Le point fort du LOUD ne réside pas seulement dans les solutions techniques qu'il peut offrir aux institutions, que ce soit l'amélioration de l'interopérabilité ou ces capacités d'annotation, mais bien dans sa capacité de résoudre des questions fondamentales communautaires et de briser les silos. C'est donc d'une approche qui est autant technique que sociale où le seuil de participation est de soumettre ses cas d'utilisation aux communautés respectives. Il s'agit ici bien des communautés IIIF et celle de Linked Art. Quant au WADM, publié sous l'égide du W3C et qui devrait rester stable pour un certain nombre d'années, il est en effet plus compliqué de le modifier ou de pouvoir directement communiquer avec cette instance, mais au cas où il y aurait une modification de celui-ci, il est assez évident que IIIF adaptera à nouveau ses spécifications pour y être conforme [17].

17. A savoir qu'une extension des spécifications est toujours permise tant que celle-ci ne devient pas contradictoire avec le standard.

L'adoption de ce socle technologique, autant par les institutions du patrimoine culturel que par les HN, permettra de parler un langage commun, de pouvoir échanger et visualiser les données de manière à tout ce qui est décrit fera parti du web et ne sera ainsi pas seulement de simples pièces échangées grâce au web. Le projet PIA s'efforcera d'ici fin janvier 2025 d'apporter des éléments de réponse et si la participation se fera principalement par l'interface utilisateur, les standards LOUD permettront de plus facilement extraire et réutiliser des métadonnées et données en ne passant que par les API.

Remerciements

Je tiens à remercier mon directeur de thèse et mes encadrants (PD Dr. Peter Fornaro, Prof. Dr. Walter Leimgruber et Dr. Robert Sanderson), les collègues du laboratoire d'humanités numériques (DHLab) de l'Université de Bâle et toutes les personnes qui participent au projet de recherche PIA, notamment les cinq autres doctorant·e·s (Murielle Cornut, Max Frischknecht, Birgit Huber, Fabienne Lüthi et Florian Spiess) ainsi que notre développeur (Adrian Demleitner).

PIA bénéficie du soutien du fonds national suisse (FNS) dans le cadre du programme interdisciplinaire Sinergia (https://data.snf.ch/grants/grant/193788).

Références

[1] I. Papadakis, K. Kyprianos, M. Stefanidakis, Linked Data URIs and Libraries : The Story So Far, D-Lib Magazine 21 (2015). URL : http://www.dlib.org/dlib/may15/papadakis/05papadakis.html. doi :10.1045/may2015-papadakis.

[2] N. Freire, G. Robson, J. B. Howard, H. Manguinhas, A. Isaac, Cultural heritage metadata aggregation using web technologies : IIIF, Sitemaps and Schema.org, International Journal on Digital Libraries (2018). URL : https://doi.org/10.1007/s00799-018-0259-5. doi :10.1007/s00799-018-0259-5.

[3] N. Freire, E. Meijers, S. d. Valk, R. Voorburg, A. Isaac, R. Cornelissen, Aggregation of linked data : A case study in the cultural heritage domain, in : 2018 IEEE International Conference on Big Data (Big Data), 2018, pp. 522–527. URL : https://doi.org/10.1109/BigData.2018.8622348. doi :10.1109/BigData.2018.8622348.

[4] N. Thalhath, M. Nagamori, T. Sakaguchi, S. Sugimoto, Wikidata Centric Vocabularies and URIs for Linking Data in Semantic Web Driven Digital Curation, in : E. Garoufallou, M.-A. Ovalle-Perandones (Eds.), Metadata and Semantic Research, Communications in Computer and Information Science, Springer International Publishing, 2021, pp. 336–344. doi :10.1007/978-3-030-71903-6_31.

[5] A. Chardonnens, La gestion des données d'autorité archivistiques dans le cadre du Web de données, 2020. URL : https://hdl.handle.net/2013/ULB-DIPOT:oai:dipot.ulb.ac.be:2013/315804, publisher : Université libre de Bruxelles.

[6] S. Münster, F. I. Apollonio, P. Bell, P. Kuroczynski, I. Di Lenardo, F. Rinaudo, R. Tamborrino, Digital Cultural Heritage meets Digital Humanities, The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences XLII-2/W15

(2019) 813–820. URL : https://www.int-arch-photogramm-remote-sens-spatial-inf-sci.net/XLII-2-W15/813/2019/. doi :10.5194/isprs-archives-XLII-2-W15-813-2019.

[7] E. Hyvönen, Using the Semantic Web in digital humanities : Shift from data publishing to data-analysis and serendipitous knowledge discovery, Semantic Web 11 (2020) 187–193. URL : https://content.iospress.com/articles/semantic-web/sw190386. doi :10.3233/SW-190386, publisher : IOS Press.

[8] R. Sanderson, The Importance of being LOUD, 2020. URL : https://www.slideshare.net/azaroth42/the-importance-of-being-loud, LODLAM 2020.

[9] D. Newbury, LOUD : Linked Open Usable Data and linked.art, in : 2018 CIDOC Conference, 2018. URL : https://cidoc.mini.icom.museum/wp-content/uploads/sites/6/2021/03/CIDOC2018_paper_153.pdf.

[10] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, J. Bouwman, A. J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C. T. Evelo, R. Finkers, A. Gonzalez-Beltran, A. J. Gray, P. Groth, C. Goble, J. S. Grethe, J. Heringa, P. A. 't Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S. J. Lusher, M. E. Martone, A. Mons, A. L. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. van Schaik, S.-A. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M. A. Swertz, M. Thompson, J. van der Lei, E. van Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao, B. Mons, The FAIR Guiding Principles for scientific data management and stewardship, Scientific Data 3 (2016) 160018. URL : https://www.nature.com/articles/sdata201618. doi :10.1038/sdata.2016.18.

[11] R. Sanderson, Cultural Heritage Research Data Ecosystem, 2020. URL : https://www.slideshare.net/azaroth42/sanderson-cni-2020-keynote-cultural-heritage-research-data-ecosystem, CNI Spring 2020 Virtual Meeting.

[12] S. Snydman, R. Sanderson, T. Cramer, The International Image Interoperability Framework (IIIF) : A community & technology approach for web-based images, in : Archiving Conference, volume 2015, IS&T, 2015, pp. 16–21. URL : https://purl.stanford.edu/df650pk4327.

[13] R. Robineau, Introduction aux protocoles IIIF, 2019. URL : https://doi.org/10.5281/zenodo.3760306.

[14] R. Sanderson, B. Albritton, R. Schwemmer, H. Van de Sompel, SharedCanvas : a collaborative model for medieval manuscript layout dissemination, in : Proceedings of the 11th annual international ACM/IEEE joint conference on Digital libraries, JCDL '11, Association for Computing Machinery, 2011, pp. 175–184. URL : https://doi.org/10.1145/1998076.1998111. doi :10.1145/1998076.1998111.

[15] M. B. Reusser, e-codices : 15 Jahre – eine Erfolgsgeschichte, ICOMOS – Hefte des Deutschen Nationalkomitees 77 (2021) 64–76 Seiten. URL : https://journals.ub.uni-heidelberg.de/index.php/icomoshefte/article/view/85149. doi :10.11588/IH.2020.1.85149.

[16] J. v. Zundert, On Not Writing a Review about Mirador : Mirador, IIIF, and the Epistemological Gains of Distributed Digital Scholarly Resources, Digital Medievalist 11 (2018) 5. URL : http://journal.digitalmedievalist.org//articles/10.16995/dm.78/. doi :10.16995/dm.78.

[17] S. Cossu, IIIF at the getty : Vision & tactics, 2020. URL : https://www.cni.org/wp-content/uploads/2020/03/scossu_iiif_getty_vision_and_tactics_cni_spring_2020.pdf, CNI Spring 2020.

[18] R. Sanderson, Discovery of IIIF resources, 2018. URL : https://www.slideshare.net/azaroth42/iiif-discovery-walkthrough, 2018 IIIF Conference.
[19] M. Appleby, D. Childress, T. Crane, J. Mixter, R. Sanderson, S. Warner, M. Whitaker, IIIF Content State API 1.0, 2022. URL : https://iiif.io/api/content-state/1.0/.
[20] N. Freire, E. Meijers, S. de Valk, J. A. Raemy, A. Isaac, Metadata aggregation via linked data : Results of the europeana common culture project, in : E. Garoufallou, M.-A. Ovalle-Perandones (Eds.), Metadata and Semantic Research, Communications in Computer and Information Science, Springer International Publishing, 2021, pp. 383–394. doi :10.1007/978-3-030-71903-6_35.
[21] R. Sanderson, P. Ciccarese, H. Van de Sompel, Designing the W3C open annotation data model, in : Proceedings of the 5th Annual ACM Web Science Conference, WebSci '13, Association for Computing Machinery, 2013, pp. 366–375. URL : https://doi.org/10.1145/2464464.2464474. doi :10.1145/2464464.2464474.
[22] J. A. Raemy, Applying Effective Data Modelling Approaches for the Creation of a Participatory Archive Platform, 2021. URL : https://infoscience.epfl.ch/record/291219.
[23] E. Delmas-Glass, R. Sanderson, Fostering a community of PHAROS scholars through the adoption of open standards, Art Libraries Journal 45 (2020) 19–23. URL : https://doi.org/10.1017/alj.2019.32. doi :10.1017/alj.2019.32.
[24] B. Heinisch, K. Oswald, M. Weißpflug, S. Shuttleworth, G. Belknap, Citizen Humanities, in : K. Vohland, A. Land-Zandstra, L. Ceccaroni, R. Lemmens, J. Perelló, M. Ponti, R. Samson, K. Wagenknecht (Eds.), The Science of Citizen Science, Springer International Publishing, 2021, pp. 97–118. URL : https://doi.org/10.1007/978-3-030-58278-4_6. doi :10.1007/978-3-030-58278-4_6.
[25] A. Demleitner, J. A. Raemy, PIA data model, 2021. URL : https://github.com/Participatory-Image-Archives/pia-data-model. doi :10.5281/zenodo.5142605.
[26] M. Appleby, T. Crane, R. Sanderson, J. Stroop, S. Warner, IIIF image API 3.0, 2020. URL : https://iiif.io/api/image/3.0/.
[27] M. Appleby, T. Crane, R. Sanderson, J. Stroop, S. Warner, IIIF presentation API 3.0, 2020. URL : https://iiif.io/api/presentation/3.0/.
[28] S. Fraser, An introduction to the Madoc Platform, 2019. URL : https://medium.com/digirati-ch/an-introduction-to-the-madoc-platform-af516a67e0.
[29] S. Blickhan, Fun with IIIF, 2022. URL : https://blog.zooniverse.org/2022/04/20/fun-with-iiif/.
[30] B. Latour, Reassembling the social : an introduction to actor-network-theory, Clarendon lectures in management studies, Oxford University Press, 2005.