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]

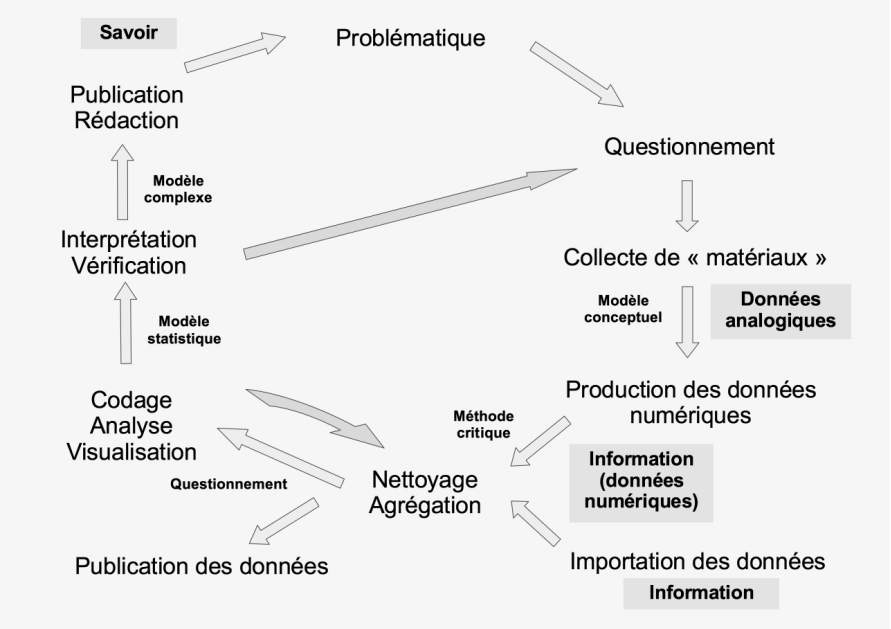

Figure 1 : Cycle de la connaissance en sciences historiques.

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,

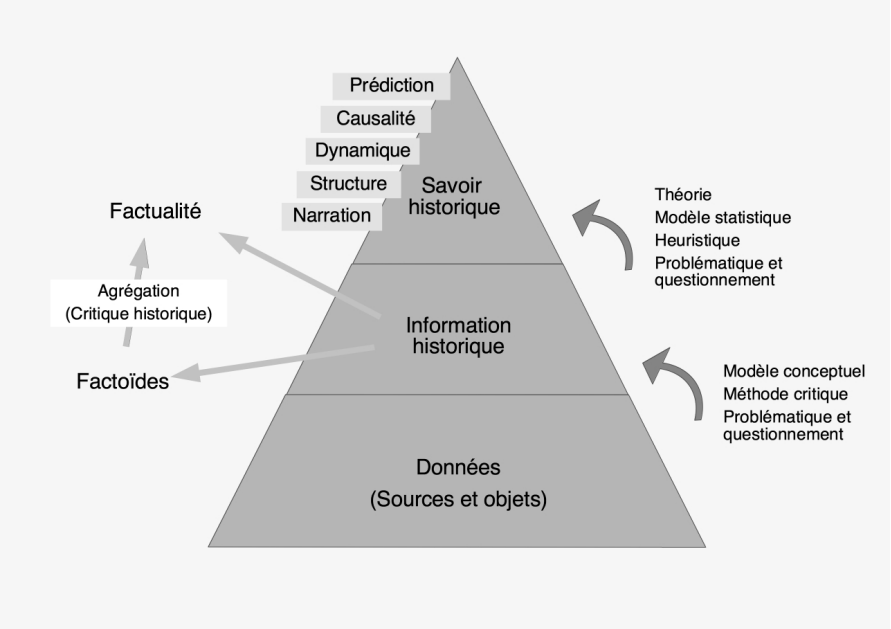

Figure 2 : Pyramide « sources, information, savoir » en sciences historiques.

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

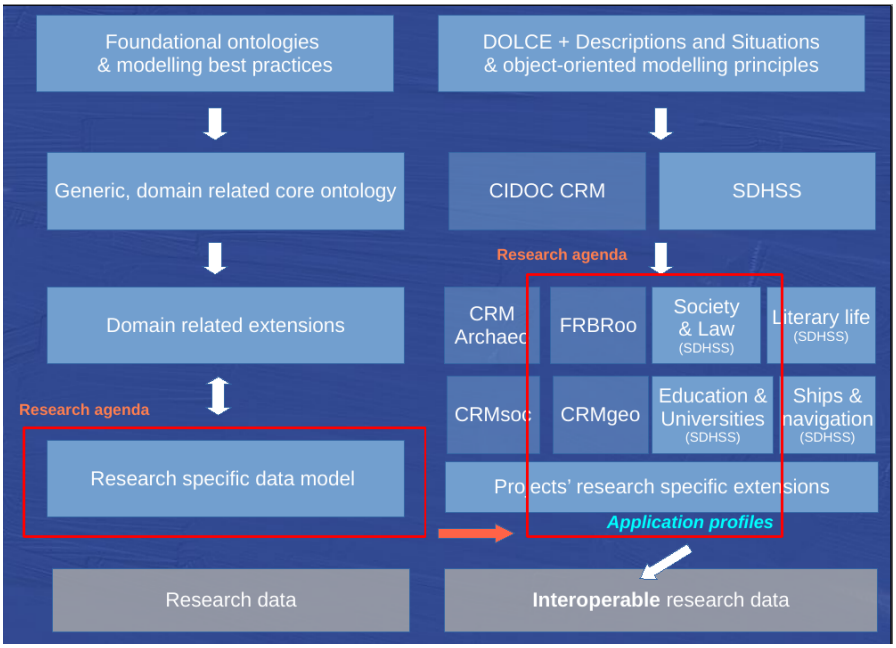

Figure 3 : Niveaux d'abstraction de la conceptualisation.

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.