Données liées ouvertes et référentiels public : un changement de paradigme pour la recherche en sciences humaines et sociales

Francesco Beretta

Données liées ouvertes et référentiels publics : un changement de paradigme pour la recherche en sciences humaines et sociales

Francesco Beretta (historien, spécialiste en systèmes d'information pour les sciences humaines et sociales, chargé de recherche au CNRS, UMR 5190 LARHRA, Lyon)

Un article publié dans *Arabesques* en 2017 faisait état d'un premier alignement avec IdRef de personnes recensées dans la plateforme *symogih.org*, un environnement virtuel de recherche (EVR) mis en place au Laboratoire de recherche historique Rhône-Alpes (LARHRA) en 2008 : "l'intégration des autorités SyMoGIH avec les IdRef doit faciliter l'ouverture de notre entrepôt vers d'autres réservoirs de qualité, tout en enrichissant les IdRef"**[1]**. Sept ans après, ce projet a connu des développements importants qui s'inscrivent dans une collaboration entre le laboratoire LARHRA et l'ABES formalisée en 2019 par une convention de coopération scientifique.

Deux éléments principaux sont au cœur de cette démarche : d'une part, la publication avec les technologies sémantiques de données de la recherche afin de faciliter leur réutilisation ; d'autre part, l'enrichissement du référentiel IdRef avec les informations issues de la recherche. La finalité de cette opération est d'encourager la réutilisation des données pour de nouvelles recherches en sciences humaines et sociales (SHS), en application des principes FAIR *(Findable, Accessible, Interoperable, Reusable).* Pourquoi est-il essentiel, dans ce contexte, de pouvoir se référer à des autorités telles celles d'IdRef?

Selon une intuition qui était à l'origine du projet *symogih.org*, il est indispensable en vue de la réutilisation des données de distinguer entre les *questions de recherche* d'un projet et *l'information* collectée pour y répondre**[2]**. Si, en effet, le *savoir* issu de la démarche scientifique peut être défini comme une *interprétation* du monde, un modèle qui répond aux questions des chercheur-ses, l'*information collectée* pour produire ce savoir doit viser une *représentation* le plus factuelle possible du monde étudié, c'est-à-dire des objets qui le composent, de leurs propriétés et de leurs relations**[3]**. Cette distinction permet de produire des données qu'on pourra réutiliser pour répondre à de nouveaux questionnements.

Grâce au Web sémantique, il devient possible de créer un graphe géant de relations entre objets du discours scientifique, relations sémantiquement explicites, et de capitaliser ainsi l'information produite par chaque projet en permettant sa réutilisation pour de nouvelles recherches. La condition est l'identification précise des objets grâce aux référentiels. Si Google a su réaliser un Giant Knowledge Graph comportant, en mars 2023, 8 milliards d'objets identifiés et 800 milliards de 'faits' (source : Wikipedia), pourquoi les SHS n'en feraient pas autant, notamment en utilisant IdRef ?

Pour que ce projet scientifique et technologique aboutisse, trois composantes sont indispensables : un référentiel partagé (1) permettant d'identifier clairement les objets du monde (personnes, organisations, concepts, etc.) ; une méthode de modélisation des relations entre objets (2) capable d'intégrer les approches de différentes disciplines ; une infrastructure distribuée durable (3) (cf. l'illustration), permettant de soutenir la démarche de recherche et l'interconnexion des données existantes.

Le référentiel IdRef (1) se prête bien à cette fin car il est connecté avec la bibliographie du SUDOC, ainsi qu'avec la plateforme Persée, les archives dans Calames ou encore l'entrepôt de publications *scienceplus.abes.fr***[4].** Il peut servir comme l'un des pivots de l'identification des objets du discours scientifique : non seulement il fait le lien vers d'autres référentiels tel celui de la BNF ou Wikidata, mais il admet un enrichissement par les chercheur-ses (soumis à un contrôle de qualité) et, en retour, il tire profit d'un processus de désambiguïsation collectif.

Il faut ensuite disposer d'une ontologie (2), c'est-à-dire d'un modèle conceptuel formalisé et partagé, modulaire et ouvert aux différentes disciplines scientifiques. Pour répondre à ce défi, le LARHRA a travaillé, sur le plan pratique, à la mise en ligne d'une application de gestion collaborative d'ontologies, OntoME (*ontome.net*). Cette plateforme permet d'étendre les standards, tel le CIDOC CRM, afin de disposer de classes et propriétés qui correspondent aux besoins des différentes disciplines SHS, et de gérer des profils applicatifs qui facilitent l'appropriation du modèle par les chercheur-ses**[5]**.

Sur le plan scientifique, l'utilisation de méthodologies de développement d'ontologies telle OntoClean, ainsi que l'analyse fondationnelle à l'aide de DOLCE, a permis de mettre en place un écosystème d'extensions du CIDOC CRM dans le projet *Semantic Data for Humanities and Social Sciences* (SDHSS)**[6].** Cette méthodologie facilite également l'intégration d'autres standards, tels *Records in Contexts* (RiC) ou le *IFLA Library Reference Model* (LRM). À noter que l'écosystème d'ontologies SDHSS se limite à proposer un

ensemble cohérent de classes et propriétés, afin de disposer d'un langage commun pour décrire les éléments essentiels de la vie sociale (le fait d'être propriétaire d'un objet, ou d'avoir un rôle dans une organisation, etc.), tandis que la gestion de vocabulaires contrôlés de types d'objets, ou de rôles sociaux, sont librement gérés par les chercheur-ses dans leurs projets respectifs, si possible en lien avec un référentiel comme IdRef.

Au niveau de l'infrastructure (3), un contrat de transfert de savoir-faire entre le CNRS et l'entreprise KleioLab a permis de créer un nouvel EVR, *geovistory.org*, qui remplace celui du projet *symogih.org* et intègre la plateforme *ontome.net*. Depuis cette année, le projet *LOD4HSS***[7]**, piloté par Tobias Hodel (professeur d'humanités numériques à l'Université de Berne), vise à promouvoir la pérennisation de cette infrastructure, qui sera portée par un consortium international d'organismes publics, et à développer de nouvelles fonctionnalités, telle l'intégration avec les graphes sémantiques de documents au format XML, encodés selon les standards TEI ou EAD. IdRef s'inscrit dans cette vision d'avenir, notamment via l'enrichissement des notices d'autorité avec des informations issues de la plateforme *geovistory.org*.

Pour les chercheurs, l'utilisation de cet EVR permettrait d'éviter deux écueils majeurs. D'une part, le fonctionnement en silos, selon le principe "nouveau projet=nouvelle base de données", qui est problématique en raison du caractère temporaire des projets et qui conduit souvent à la disparition des plateformes, et des données, une fois les financements terminés. D'autre part, l'absence d'une sémantique commune rend la réutilisation des données difficile voire impossible. Même en se servant du même outil (que ce soit Heurist, NodeGoat ou Wikibase) les données restent 'prisonnières' de dépôts étanches les uns aux autres et leur interopérabilité est mise à mal par des choix de modèles conceptuels divergents ou contradictoires**[8]**.

Certes, des méthodologies existent pour transformer ces données et les aligner avec les référentiels et une ontologie partagée. Un projet pilote a été mené dans le cadre de la collaboration entre l'ABES et le LARHRA, dans le contexte de l'ANR HisArc-RDF, qui a permis de créer un prototype de processus de transformation et publication de données sous forme de données liées ouvertes (Linked Open Data, LOD)**[9]**: après alignement avec les IdRef et en utilisant le standard FRBRoo de l'IFLA, une partie des données du projet PRELIB, consacré au monde littéraire breton, est désormais accessibles sur le serveur SPARQL du projet *dataforhumanities.org***[10]**. Reste que cette démarche comporte des coûts supplémentaires, rarement prévus dans le budget des projets.

L’évolution vers la publication de données de la recherche sous forme de LOD alignés avec les référentiels (si possible produits dès l’origine comme tels) permet d’envisager un renouvellement important des SHS grâce à un changement d’échelle du volume d’information disponible, virtuellement infini et de bonne qualité, facilement réutilisable grâce aux technologies du web sémantique. Le potentiel est tel qu’on peut prévoir un changement de paradigme dans ces disciplines, une transformation de leur manière de produire le savoir et de former les nouvelles générations de chercheur-ses**[11]**. Pour ce faire, une infrastructure collaborative et ouverte telle *geovistory.org*, capable d’accueillir grâce aux méthodologies sémantiques une grande variété de projets en SHS, par exemple de type Collex-Persée, est indispensable. De même en va-t-il de l’intégration des compétences liées aux LOD dans les métiers des Bibliothèques, de l’Information et du Patrimoine, afin d’accompagner les chercheur-ses, et le public, dans la transition numérique.

---

[1]Pierre Vernus, « SyMoGIH, de l’UMR 5190 – Larhra, et les ‘objets historiques’ », *Arabesques*, 85 | 2017, 14.

[2]Francesco Beretta and Pierre Vernus, « Le projet SyMoGIH et la modélisation de l’information : une opération scientifique au service de l’histoire », *Les Carnets du LARHRA*, 1, 2012, 81–107.

[3]Tom Gruber, « Ontology », in Liu, Ling, and M. Tamer Özsu, éds., *Encyclopedia of Database Systems, Second Edition* (Springer, 2018), 2574–76 https://doi.org/10.1007/978-1-4614-8265-9

[4]Yann Nicolas, « Scienceplus.abes.fr : une nouvelle base de données au service de la science ouverte », *Arabesques*, 103 | 2021, 22.

[5]Francesco Beretta, « A Challenge for Historical Research: Making Data FAIR Using a Collaborative Ontology Management Environment (OntoME) », Semantic Web, 12.2 (2021), 279–94, https://doi.org/10.3233/SW-200416

[6]Id., « Interopérabilité des données de la recherche et ontologies fondationnelles : un écosystème d’extensions du CIDOC CRM pour les sciences humaines et sociales », in

Nicolas Lasolle, Olivier Bruneau, and Jean Lieber, éds, *Actes des journées Humanités Numériques et Web sémantique*, (Nancy, France, 2022), pp. 2–22 <https://doi.org/10.5281/zenodo.7014341>

[7]https://www.geovistory.org/lod4hss

[8]https://www.mediawiki.org/wiki/Wikibase/FAQ: “Wikibase users can design their own data model. Are there downsides to this?”

[9]https://dataforhumanities.org/sparql-endpoint/prelib-v1 .

[10]François Mistral, « Des catalogues de bibliothèques aux projets en humanités numériques : les autorités IdRef font le lien », *Arabesques*, 105 | 2022, 16-17.

[11]Francesco Beretta, « Données ouvertes liées et recherche historique : un changement de paradigme », *Humanités numériques*, 7, 2023, https://doi.org/10.4000/revuehn.3349