humanités numériques

Humanités numériques

7 | 2023
Numérisation du patrimoine et modélisation des connaissances

Données ouvertes liées et recherche historique : un changement de paradigme

Linked Open Data and Historical Research: a Paradigm Shift

Francesco Beretta

OpenEdition Journals

Édition électronique
URL : https://journals.openedition.org/revuehn/3349
DOI : 10.4000/revuehn.3349
ISSN : 2736-2337

Éditeur
Humanistica

Référence électronique
Francesco Beretta, « Données ouvertes liées et recherche historique : un changement de paradigme », *Humanités numériques* [En ligne], 7 | 2023, mis en ligne le 01 juillet 2023, consulté le 14 juillet 2023.
URL : http://journals.openedition.org/revuehn/3349 ; DOI : https://doi.org/10.4000/revuehn.3349

humanités
numériques

7 | 2023
Numérisation du patrimoine et modélisation des connaissances
MODÉLISATION DES CONNAISSANCES

Données ouvertes liées et recherche historique : un changement de paradigme

Linked Open Data and Historical Research: a Paradigm Shift

Francesco Beretta

Résumés

Dans le contexte de la transition numérique, le Web sémantique et les données ouvertes liées (*linked open data* [LOD], en anglais) jouent un rôle de plus en plus central, car ils permettent de construire des « graphes d'information » (*knowledge graphs*, en anglais) reliant l'ensemble des ressources du Web. Ce phénomène interroge les sciences historiques et soulève la question d'un changement de paradigme. Après avoir précisé ce qu'il faut entendre par « données », l'article analyse la place qu'elles occupent dans le processus de production du savoir. Il présente les principales composantes du changement de paradigme, en particulier le potentiel des LOD et d'une sémantique robuste en tant que véhicules d'une information factuelle de qualité, intelligible et réutilisable. S'ensuit une présentation des projets d'infrastructure réalisés au sein du Laboratoire de recherche historique Rhône-Alpes (Larhra) : *symogih.org*, *ontome.net*, *geovistory.org*. Leur but est de faciliter la transition numérique grâce à un outillage construit en cohérence avec l'épistémologie des sciences historiques et de contribuer à la réalisation d'un « graphe d'information » disciplinaire.

In the context of the digital transition, the semantic Web and linked open data (LOD) play an increasingly central role as they allow the construction of knowledge graphs linking together the resources of the Web. This phenomenon questions the historical sciences and raises the major issue of a paradigm shift. After clarifying what the meaning of "data" is, the paper analyses their place in the process of knowledge production. It presents the main components of the paradigm shift, and in particular

the potential of LOD and robust semantics as vehicles for high-quality, intelligible and reusable factual information. We then present the infrastructure projects carried out at the Larhra (Laboratoire de recherche historique Rhône-Alpes) with the aim of facilitating the digital transition: *symogih.org*, *ontome.net*, *geovistory.org*. These Web applications are developed in line with the epistemology of the historical sciences and contribute to the realisation of a disciplinary "information graph".

Entrées d'index

MOTS-CLÉS : histoire, épistémologie, Web sémantique, interopérabilité, modélisation des connaissances

KEYWORDS: history, epistemology, semantic Web, interoperability, knowledge modelling

Introduction

1 La recherche historique s'inscrit aujourd'hui dans le contexte d'une évolution de la société dans laquelle le numérique envahit tous les domaines, de la production industrielle à la vie privée, des activités récréatives aux diagnostics médicaux. Au cœur de ce phénomène se trouvent les « données », qu'il s'agisse de *fake news* ou d'informations de qualité. « *Data is the new oil* », affirme-t-on souvent, et la croissance exponentielle du volume de données, en particulier dans les plateformes des GAFAM (Google, Apple, Facebook, Amazon et Microsoft), permet certainement à ces entreprises de produire des bénéfices commerciaux gigantesques, d'influencer les choix individuels et d'orienter la vie publique. Si la transition numérique peut être comparée au passage du manuscrit à l'imprimé et à l'évolution qu'il a provoquée – au cours de trois siècles – d'une culture d'élite vers une culture de masse, la transformation à laquelle nous assistons est totalement inédite, tant par sa rapidité que par ses conséquences.

2 Dans ce contexte, le Web sémantique, avec le mouvement des données ouvertes liées (*linked open data* [LOD], en anglais), initié par Tim Berners-Lee en 2001 [1], joue un rôle de plus en plus central car il relie et donne du sens aux autres ressources du Web et permet de mobiliser d'une manière totalement nouvelle l'information que contiennent les données. On va jusqu'à parler de « graphes de connaissances », nommés ainsi d'après le *giant knowledge graph* créé par Google, contenant des centaines de milliards de propriétés au sujet de plus de cinq milliards d'entités [2], qui permet à l'entreprise d'optimiser son moteur de recherche et qui contribue à renforcer sa position dominante. Je préfère utiliser le terme « graphe d'information », en tant que spécialisation de celui de « système d'information », et réserver le terme de « connaissance » à l'information contextualisée et interprétée qui est le propre du savoir humain.

3 Le monde de la recherche historique, tout comme celui de la conservation des biens culturels, doit s'interroger et se positionner par rapport à ce phénomène à la fois technologique et social. Depuis quelques décennies, on assiste au développement de systèmes d'information de plus en plus sophistiqués : il n'y a qu'à penser aux ressources des bibliothèques numériques disponibles sur Europeana [3] ou aux données ouvertes mises à disposition par la Bibliothèque nationale de France dans BNF Data [4], qui sont reliées avec l'ensemble des principales bibliothèques mondiales par l'intermédiaire du fichier d'autorités VIAF [5]. Du côté de la recherche en sciences humaines et sociales, l'Agence bibliographique de l'enseignement supérieur (ABES) publie son catalogue Sudoc (Système universitaire de documentation [6]) et ses notices d'autorité IdRef [7] sous forme de LOD interrogeables, tandis que l'infrastructure de recherche Huma-Num a créé des plateformes d'agrégation et de publication des données de la recherche [8]. Cette dynamique s'inscrit dans la stratégie des agences publiques de la recherche qui encouragent, voire obligent, les producteurs à ouvrir leurs données. Pour la France, on peut mentionner OpenData France [9] et la Feuille de route pour la science ouverte du CNRS [10].

4 Dans le contexte de la transition numérique, les principes FAIR (*findable*, *accessible*, *interoperable*, *reusable*) ont été formulés afin de guider la production et la publication des données issues de la recherche pour qu'elles soient « plus facilement accessibles, comprises, échangeables et réutilisables [11] ». Leur finalité est de promouvoir les bonnes pratiques permettant la réutilisation des données en vue de répondre à de nouveaux questionnements [12]. Les chercheurs et chercheuses sont ainsi invités à publier non seulement leurs résultats sous forme d'articles ou de livres, mais encore à mettre à disposition les données elles-mêmes ayant servi à les établir [13]. Il s'agit, d'une part, de garantir la reproductibilité des résultats et, d'autre part, de permettre à d'autres chercheurs et agents du monde économique ou culturel de profiter de ces gisements d'information de qualité, produits grâce aux financements publics. Les LOD jouent un rôle central dans ce processus car il s'agit de données structurées reliant les différentes ressources du Web, publiées dans un format standardisé et donc actionnables aussi bien par les humains que par les machines.

5 Lorsque les LOD sont produites en utilisant une sémantique formalisée et partagée qui en explicite le sens – « *data use a formal, accessible, shared, and broadly applicable language for knowledge representation* » (principes FAIR, I1) –, elles peuvent servir de fondement à la réalisation de *knowledge graphs* disciplinaires et faciliter ainsi la réutilisation de l'information de qualité, dont les données scientifiques sont le support numérique. Dans un article à la fois critique des pratiques courantes et très stimulant, Giancarlo Guizzardi souligne l'importance des *ontologies de domaine* comme *meaning contracts* qui, en explicitant et formalisant la signification des données, les rendent intelligibles et réutilisables. À condition toutefois d'appliquer à leur développement les méthodologies de l'*ontologie formelle* : cette discipline scientifique a développé les principes indispensables à la production d'information sémantiquement interopérable, notamment la méthodologie OntoClean (Guarino et Welty 2009), et produit les ontologies fondationnelles (*upper ontologies* [14]) en tant qu'instruments de vérification et d'intégration conceptuelle des ontologies de domaine (Guizzardi 2020).

6 Devant l'ampleur de ce phénomène et son impact sur la mise à disposition de données de la recherche réutilisables, une interrogation surgit inévitablement : est-ce que nous sommes en présence d'un changement de paradigme ? Pour répondre, je ne vais pas dresser un bilan, qui ne peut être réalisé que collectivement (Genet 2011 ; Meroño-Peñuela *et al.* 2015), mais je proposerai quelques réflexions concernant l'impact de la transformation numérique sur le processus de connaissance en sciences historiques. Ces réflexions résultent d'une quinzaine d'années d'expérience dans la mise en place de systèmes d'information collaboratifs pour la recherche, d'enseignement en méthodologies numériques (niveaux bachelor et master) et d'accompagnement de multiples recherches doctorales et projets de recherche collectifs. Dans la suite de l'article, je présenterai tout d'abord un exemple illustrant la transformation numérique, tout en précisant ce qu'il faut entendre par « données ». J'analyserai ensuite la place qu'elles occupent dans le processus de production du savoir et leur articulation avec l'information factuelle, socle d'une connaissance historique scientifiquement fondée. Puis je mettrai en évidence les principales composantes du changement en cours, ainsi que le rôle central que jouent les LOD dans ce processus. Enfin, je présenterai le projet d'infrastructure réalisé dans le cadre de mon activité de chercheur au Centre national de la recherche scientifique (CNRS), ayant pour but de faciliter la transition numérique grâce à un outillage construit en cohérence avec l'approche épistémologique présentée et favorisant la production collaborative et cumulative d'information.

Un changement de paradigme pour les sciences historiques ?

7 Le concept de paradigme a été utilisé par Thomas Kuhn en 1962 dans son ouvrage *La Structure des révolutions scientifiques* pour décrire la structure intellectuelle des disciplines et analyser les ruptures qui amènent aux révolutions scientifiques (Kuhn 1996 [1962]). Retenons deux éléments essentiels de ce terme polysémique, qui a suscité de nombreuses discussions et est utilisé en épistémologie, en histoire des sciences et en sociologie de l'activité scientifique : d'une part, le paradigme est constitué par l'ensemble des notions, méthodologies, pratiques et acquis communs qui fondent et structurent une communauté disciplinaire ; d'autre part, dans son sens issu de l'Antiquité, il comprend les pratiques pédagogiques appliquées au cours des formations universitaires dans le but de permettre l'apprentissage des compétences indispensables à l'exercice d'une discipline scientifique. Étant donné que la finalité de la connaissance scientifique est la production du savoir, le paradigme permet aux étudiants d'apprendre les méthodes et règles légitimes à l'intérieur d'une communauté disciplinaire.

8 Poser la question d'un changement de paradigme revient donc à s'interroger sur la *transformation de l'activité de production de savoir et d'apprentissage disciplinaire en sciences historiques dans le nouveau contexte numérique* : quel est l'impact de la croissance exponentielle des données numériques disponibles, et du renouvellement méthodologique et pédagogique qui en découle, sur la recherche historique et ses résultats ? Qu'il s'agisse d'une vraie rupture ou d'une évolution très rapide et d'extrême

ampleur, force est de constater que l'activité de recherche s'est considérablement transformée au cours des vingt dernières années du fait du remplacement du papier et des pratiques analogiques par les données, logiciels et infrastructures numériques.

9 Mais en quoi consistent donc précisément les données numériques, quel est leur statut épistémique dans le contexte de la recherche en sciences historiques ? Je répondrai à cette question à partir d'un exemple lié à l'histoire de Corse, afin d'illustrer concrètement les enjeux [15]. Le site Web *OpenDataCorsica* publie une « Liste des monuments historiques en Corse [16] ». Celle-ci regroupe plus de trois cents monuments avec quelques éléments d'identification assortis de leur géolocalisation, ce qui permet de les placer automatiquement sur une carte. Si l'on télécharge cette liste et produit une distribution des monuments par type, on constate que les monuments les plus fréquents sont de type « architecture religieuse » (164), suivi par les types « architecture militaire » (49), « domestique » (36) puis « site archéologique » (26). Ces données sont issues de la base « Patrimoine architectural (Mérimée) » du ministère de la Culture, qui fournit des renseignements supplémentaires avec des photos de chaque monument, par exemple le site « Alignement de menhirs de Pagliajo [17] ». En utilisant l'identifiant Mérimée (PA00099118), on peut exécuter une requête SPARQL [18] sur la plateforme *Wikidata* [19] (figure 1) et retrouver l'identifiant VIAF du site du Pagliajo ou Palaghju (Sartène), de là rebondir sur la page *Wikipédia* du site mégalithique [20], qui contient des images et informations supplémentaires sous forme de textes, ou sur la version sous forme de données DBPedia [21], qui relie ce site à d'autres pages Web et ressources. On peut rebondir depuis DBPedia vers la notice de Roger Grosjean (1920-1975 [22]), archéologue et directeur du Centre de préhistoire corse (1964 [23]), et, par l'intermédiaire du fichier d'autorités IdRef [24], atterrir sur la version numérique de son article « Les alignements de Pagliaiu (Sartène, Corse) », publiée sur la plateforme Persée (Grosjean 1972).

Figure 1. Extrait du graphe *Wikidata* concernant les sites archéologiques corses (requête de la note 19) avec le site du Pagliajo

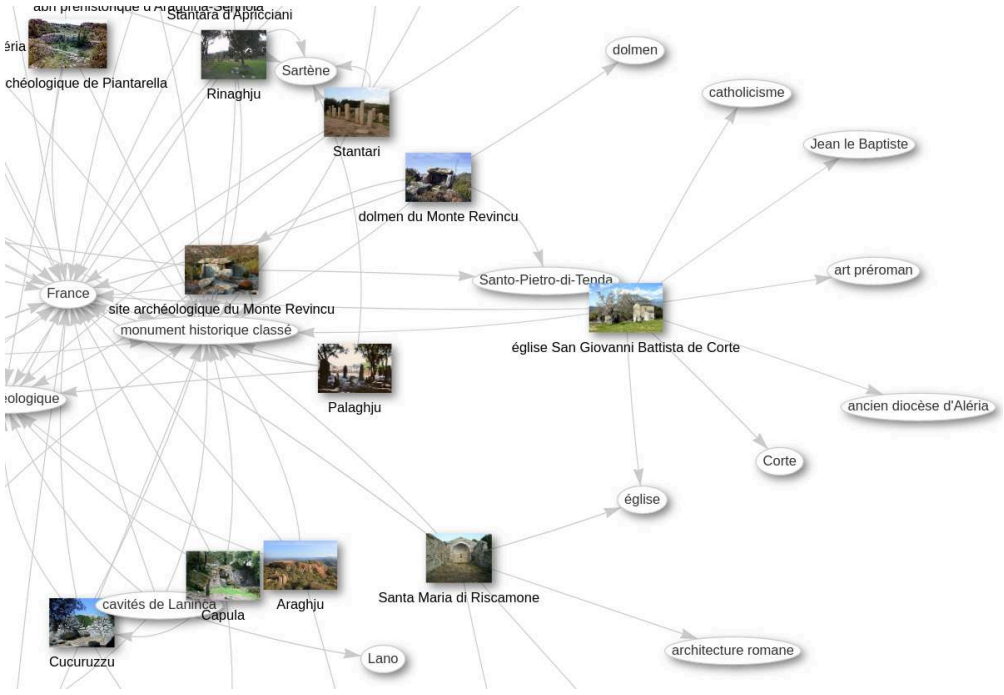

Image produite par l'auteur

10 À noter que si on est un peu « bricoleur », on peut facilement, en utilisant les langages R ou Python et en appliquant les recettes présentées sur le site *Programming Historian* [25], automatiser ce processus et récupérer tous les textes, données et images de l'ensemble des monuments historiques de Corse ou, si l'on préfère, uniquement ceux des sites archéologiques, à partir d'un questionnement précis, puis rebondir sur d'autres sites en Sardaigne ou sur le continent. Si on dispose du savoir-faire nécessaire, on peut donc, en une heure, se procurer un riche ensemble d'images, d'informations et de travaux concernant les sites mégalithiques, afin de les utiliser pour ses propres recherches ou, pourquoi pas, pour alimenter une application de découverte de la région pour *smartphone*, alors qu'il aurait fallu auparavant une semaine de travail, dans une ou plusieurs bibliothèques, pour disposer du même lot d'informations.

11 À noter aussi que l'on ne parle pas d'un seul site, mais de 30, 300, voire de 3 000 si les données sont disponibles. L'accès à l'information change donc radicalement d'échelle tant par son volume que par sa richesse et sa complexité. L'efficacité de ce processus de recherche est fondée sur les LOD : ces graphes porteurs d'information consistent en arcs ou propriétés, dont la sémantique est prédéfinie, qui relient un ensemble de nœuds ou sommets (figure 1), connectés à leur tour avec les sommets d'autres gisements représentant les mêmes entités du monde. Grâce aux infrastructures LOD de DBPedia, *Wikidata*, IdRef ou VIAF, on a ainsi pu repérer, par rebondissements successifs, un ensemble de ressources sur lesquelles on pourra réunir, en quelques requêtes, une abondante bibliographie issue des points d'accès SPARQL du Sudoc et de la BNF.

12 Les données numériques apparaissent ainsi comme le véhicule d'un graphe d'information géant (*giant knowledge graph*), dont l'accès et l'analyse sont prodigieusement facilités. La même information pourrait se trouver, comme autrefois, sur un support papier et en serait alors pour ainsi dire « prisonnière ». La transformation en LOD permet de rendre l'information « actionnable », de la filtrer, de l'interroger dans son ensemble, de l'interconnecter et de l'enrichir avec d'autres ressources du Web puis de l'analyser de manière totalement nouvelle. Si les données sont le support de l'information, celle-ci peut être définie comme *représentation des objets du monde*, qu'ils soient physiques ou conceptuels (personnes, organisations, textes, lieux, etc.), ainsi que *de leurs propriétés et de leurs relations*.

13 Concernant l'exemple de la préhistoire corse, on dispose d'abord des images d'objets, c'est-à-dire des *données non structurées* comme les photographies, fichiers audio ou numérisations de travaux de recherche. L'information est dans ce cas une représentation des objets à un moment donné, selon un angle de vue précis ou une méthode de transcription particulière. Même dans le cas de reproductions d'objets, les données numériques ne sont jamais « données », disponibles pour ainsi dire « naturellement ». Elles sont toujours *construites,* puisqu'il s'agit du support d'une information spécifique définie et conceptualisée en fonction des besoins des chercheurs. On a ensuite l'information saisie sous forme de tableur : dans ces *données semi-structurées*, telle la « Liste des monuments historiques en Corse », les propriétés des objets sont représentées sous forme de chiffres ou de chaînes de caractères. Les textes transcrits au format XML, selon les directives de la Text Encoding Initiative, re-

lèvent également de cette catégorie. L'information contenue dans ces données est certes utilisable pour la recherche, mais il n'est pas possible de la relier explicitement à d'autres ressources (Beretta 2016).

14 Si on veut représenter les relations entre objets, il existe plusieurs technologies, allant des bases de données relationnelles aux bases de données orientées graphe et aux technologies sémantiques, telles *Wikidata* ou BNF Data, qui stockent l'information sous forme de *données structurées.* Dans ce cas, l'information et les données qui en sont le support représentent les *relations* dans l'espace et dans le temps qui subsistent entre objets identifiés. Ces relations ont été préalablement définies dans un modèle conceptuel ou une ontologie afin de les rendre compréhensibles et opérables. Si pour l'être humain, les textes ont le contenu sémantique le plus riche, pour l'ordinateur et les technologies d'intelligence artificielle, ce sont les LOD, les graphes sémantiques, qui présentent un potentiel extraordinaire en matière de traitement et d'analyse de l'information. Et ce avec une étendue et une rapidité inatteignables par le cerveau humain.

15 En utilisant aujourd'hui un tableur ou une base de données *ad hoc* pour stocker les données, non seulement on se prive de toute la richesse sémantique des LOD et de leur potentiel de traitement, mais encore on risque de ne pas pouvoir réutiliser l'information collectée. La communauté de recherche va ainsi continuer à parcourir mille fois le premier kilomètre, alors qu'une démarche collaborative de collecte de l'information, soutenue par des plateformes de recherche fondées sur les technologies sémantiques, permet de parcourir ensemble des milliers de kilomètres et de disposer, en très peu de temps et en faisant levier sur une curation collective des données [26], d'un graphe d'information de grande complexité, qualité et richesse. Certes, des plateformes telles que DBPedia, graphe de données extrait de *Wikipédia*, ou son concurrent *Wikidata*, qui résultent d'une collecte d'information par les non-spécialistes et largement automatisée, ont une grande utilité. Mais c'est seulement grâce à des *plateformes disciplinaires spécifiques* que la communauté en sciences historiques et, plus largement en sciences humaines et sociales, pourra tirer entièrement profit de la transformation numérique, afin de produire, grâce à la curation collective des données, une représentation du monde social, économique, politique et culturel du passé, dans toute sa complexité, sous forme de LOD sémantiquement intelligibles et actionnables, donc réutilisables pour répondre à de nouveaux questionnements scientifiques.

La production du savoir historique dans le contexte de la révolution numérique

16 Afin de mieux cerner les composantes du changement de paradigme en cours, je vais proposer une analyse qui articule *données numérique*s et *information* dans le contexte de la production du *savoir*. Je le ferai à l'aide de deux modèles qui résument, à partir de deux points de vue différents, le processus de connaissance en sciences historiques. Le premier modèle est issu de l'épistémologie de la connaissance en histoire et s'inspire des étapes de l'élaboration du savoir formulées par Henri-Irénée Marrou (1961) sous forme de courbe parabolique dans un travail classique consacré au « métier d'historien » (figure 2). Le choix de présenter ici ce même processus sous forme de *cycle* souligne la dimension itérative de la connaissance, qui enrichit et questionne le savoir existant. Le deuxième modèle interprète, du point de vue des sciences historiques, la représentation classique sous forme de pyramide (figure 3), issue des sciences de l'information, des différents niveaux épistémiques que représentent les données, l'information et le savoir (Rowley 2007). La *connaissance* est entendue ici comme processus, le *savoir* comme contenu et résultat de l'analyse et de l'interprétation de l'*information* : le modèle du cycle représente donc la dynamique de la connaissance tandis que la pyramide en décrit les étapes sous forme de strates épistémiques.

Figure 2. Cycle de la connaissance en sciences historiques

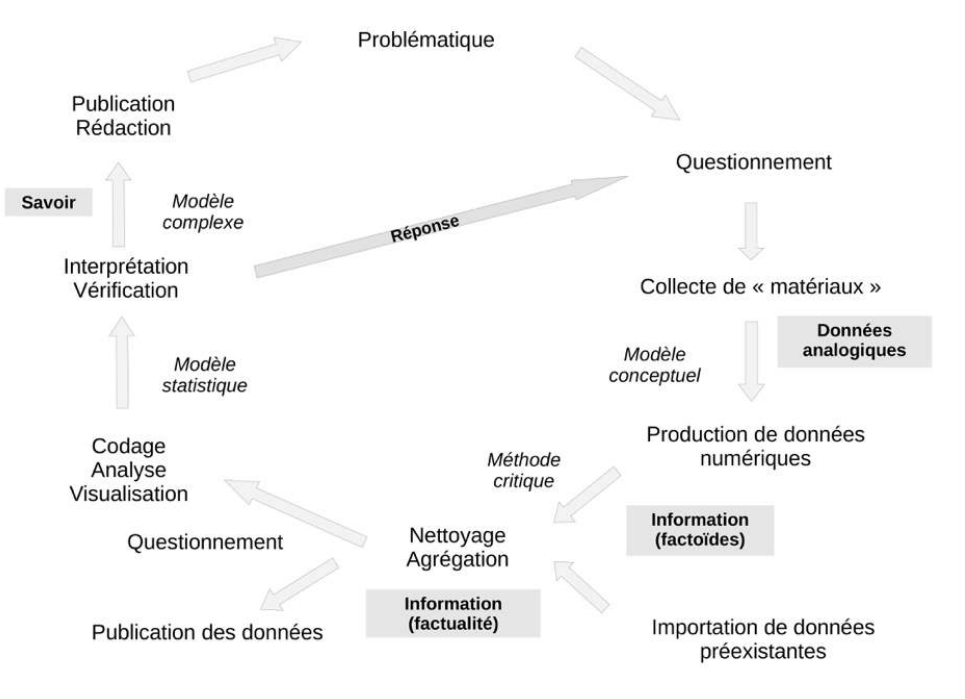

Image produite par l'auteur

17 On le sait, toute recherche doit partir d'une problématique s'inscrivant dans l'horizon du savoir existant et qui définit l'angle d'approche du sujet étudié, la méthodologie et la question centrale (figure 2). Par exemple, dans une approche d'histoire intellectuelle des sciences, on peut s'interroger sur les conditions et les dynamiques de diffusion de la doctrine du mouvement de la Terre à l'époque moderne. Cette question

générale doit être articulée avec un questionnement plus précis, concernant par exemple les carrières des astronomes et leur insertion dans les réseaux savants, en relation avec l'analyse du contenu de leurs écrits, en restreignant éventuellement l'étude à une région ou à une catégorie spécifique, par exemple les enseignants universitaires. Cette première étape est indispensable afin de pouvoir choisir ensuite les sources à utiliser et définir l'information qu'il faudra collecter pour répondre au questionnement. L'*information* est conçue dans ce contexte comme une *représentation* des propriétés et des relations des objets étudiés dans la perspective d'une factualité critiquement établie sur laquelle s'appuie la production du savoir.

18 Si on prend maintenant en considération la pyramide (figure 3), les *données* qui en forment le socle sont à entendre au sens premier et étymologique, issu du latin *datum*, qui souligne l'indépendance par rapport à l'observateur : en sciences historiques, il s'agit des *sources*, écrites ou matérielles, en tant que reflet et traces de la réalité humaine du passé. Pour mener à bien une recherche, il faut opérer un choix dans la masse des sources – les données de la pyramide – tout en adoptant un point de vue correspondant au questionnement, afin de décider quelle information sera retenue. Si les sources sont donc *données*, l'information qui en est extraite en appliquant les principes de la méthode critique est toujours *construite*, même si stockée sous forme de *données numériques*. Étant polysémique, le terme « données » doit être manié avec précaution : les données numériques n'appartiennent pas à la strate épistémique des données, mais bien à celle de l'information dont elles constituent le support informatique.

Figure 3. Pyramide « sources, information, savoir » dans le contexte des sciences historiques

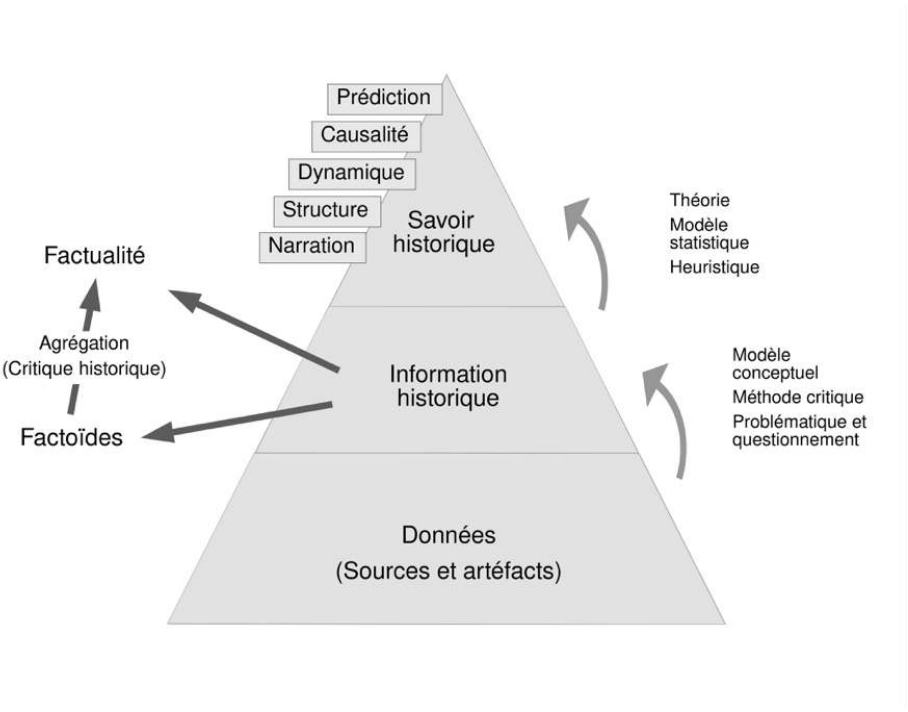

Image produite par l'auteur

19 À noter que la démarche présentée s'applique également à l'historiographie traditionnelle, qui utilise les supports analogiques, notamment les « fiches Bristol », pour collecter l'information. Quant à l'histoire nu-

mérique, la question du modèle conceptuel et du choix de la technologie de stockage à adopter se pose à ce stade : si pour un dépouillement sériel d'une seule source, une feuille de tableur bien faite peut faire l'affaire, dès que l'on mobilise plusieurs sources, ou que l'on veut observer les relations subsistant entre de multiples objets dans l'espace et dans le temps, l'utilisation d'une base de données relationnelle ou, encore mieux, d'un serveur de LOD devient indispensable.

20 Il convient également de relever que l'information dans la pyramide s'articule à deux niveaux : on peut viser une reproduction du contenu des sources, c'est-à-dire de leur manière de présenter plus ou moins fidèlement les faits, niveau qui est appelé depuis quelques années celui des *factoïdes* (Pasin et Bradley 2015) ; ou essayer de déployer toute la sagacité de la méthode critique, afin d'approximer la *factualité*, l'information représentant dans ce cas autant que possible les phénomènes du passé comme tels. Dans la plupart des recherches, l'*information factuelle* représente le socle indispensable de la production du savoir. En effet, si on veut comparer, pour revenir à notre exemple, les carrières des astronomes du XVII^e^ siècle, il ne suffit pas de se limiter à collecter les multiples mentions, dans différentes sources, des mêmes segments du parcours concernant les mêmes personnes, mais il faut les agréger afin d'identifier et de reconstituer la carrière de chaque personne. En cas de désaccord entre les sources, il sera donc inévitable de faire des choix, afin que l'analyse ne soit pas biaisée par la redondance des factoïdes et que l'information produite soit la meilleure approximation possible de la factualité. Ces choix, de même que l'origine et la qualité attribuées à l'information extraite des sources, doivent être documentés avec précision afin de permettre la réutilisation de l'information : dans le contexte des LOD, il existe des ontologies et technologies qui permettent d'ajouter à chaque donnée les métadonnées nécessaires [27]. Un approfondissement de cette question dépasse toutefois le cadre du présent article.

21 Une fois cette agrégation effectuée, la voie se sépare entre histoire analogique et histoire numérique : la première doit nécessairement passer à une phase d'interprétation, car elle ne dispose pas de la possibilité – en dehors de la lecture des notes prises – d'interroger l'information récoltée, de la rendre « actionnable » ; la deuxième peut appliquer à l'information, transformée en données numériques et opportunément nettoyée et codée, une panoplie d'outils et de logiciels statistiques, d'analyse de réseaux, de représentation et d'analyse spatiale, etc. (figure 2). C'est là qu'apparaît un premier aspect du changement de paradigme : pour l'histoire numérique, la phase d'interprétation surviendra seulement dans un deuxième temps et sera alimentée par le résultat des analyses effectuées avec différents logiciels. Certes, le *modèle* – au sens statistique – qui ressort de ces analyses a une fonction principalement heuristique, car les représentations mathématiques et visuelles qui résultent des outils numériques nécessitent toujours une discussion critique ainsi qu'une contextualisation et interprétation. Mais en même temps, ces logiciels de plus en plus performants peuvent faire apparaître des phénomènes historiques significatifs qu'il serait autrement impossible de voir « à l'œil nu » – par exemple, les outils d'analyse de séquences permettent la comparaison de configurations récurrentes dans la car-

rière de centaines d'astronomes – et ce, en tirant profit d'un volume et d'une complexité de l'information impossibles à maîtriser avec le seul cerveau humain.

22 Le cycle de connaissance aboutit à la production du *savoir* comme réponse à un questionnement. Il importe de souligner la distinction épistémique essentielle – illustrée dans le modèle de la pyramide – qui subsiste entre le *savoir* historique et l'*information* sur laquelle il se fonde : l'information est définie en tant que représentation des propriétés et relations des objets du monde, dans une perspective factuelle indépendante autant que possible de la problématique de recherche – et ce, même si sa collecte a été effectuée en fonction de celle-ci –, tandis que le savoir est compris comme *modèle du monde*, dans toute sa complexité, construit dans la perspective de répondre à un questionnement. Le savoir pourra concerner, en fonction de différentes problématiques, la reconstitution précise des événements sous forme de narration, ou l'analyse de la structuration d'un champ scientifique autour d'institutions académiques ou éditoriales, ou l'évolution dans le temps de la théorie explicative dominante et la recherche des causes de cette dynamique. Si le volume d'information est suffisant, on pourrait même proposer des hypothèses quant aux évolutions futures d'un phénomène précis, approche inimaginable à défaut d'un support numérique de l'information.

23 Il ressort de cette analyse que l'interopérabilité et la réutilisation des données numériques issues de la recherche historique concernent principalement l'information : les LOD sont un *graphe d'information* plutôt qu'un graphe de connaissances ou du savoir. Il est en effet préférable de réserver le premier terme, *connaissance*, au processus qui aboutit à la production du *savoir*, les deux se situant dans l'esprit des chercheurs et non dans les données, qui sont des artéfacts numériques impliquant vérification et interprétation. Au premier aspect du changement de paradigme présenté ci-dessus, que l'on pourrait appeler une heuristique outillée par les logiciels d'analyse des données, s'en ajoute un second, qui comporte un changement d'échelle : les nouvelles méthodologies de gestion des systèmes d'information et les LOD permettront non seulement d'interroger les données récoltées par un seul chercheur, mais encore de tirer profit d'un nombre sans cesse croissant de gisements d'information interconnectés. L'information sera donc disponible en quantité jusque-là inimaginable, munie de métadonnées concernant sa provenance et sa qualité, alignée sur les référentiels tels VIAF ou *Wikidata* dans la logique des principes FAIR, et pourra ainsi être réutilisée pour de nouvelles recherches d'une ampleur inégalée par rapport à celles précédant la transition numérique.

24 Dans la logique de la science ouverte, on pourra de cette manière construire un modèle explicatif complexe de la réalité historique et soumettre le savoir produit à une démarche de réfutabilité (ou de falsification) [28] en publiant les informations sémantiquement explicitées qui auront permis de l'établir. On inscrira ainsi la recherche historique dans une *démarche scientifique de reproductibilité*. C'est dans ce second aspect que se manifeste clairement le changement de paradigme en cours.

Les composantes du changement de paradigme

25 Ce changement de paradigme peut être articulé autour de cinq composantes : l'automatisation, l'interconnexion, la sémantique, le volume, l'analyse. Grâce aux techniques de traitement automatique du langage naturel ainsi qu'aux méthodologies d'apprentissage automatique, les ordinateurs sont en mesure de reconnaître de manière de plus en plus efficace les objets (personnes, organisations, lieux, etc.) que les corpus textuels mentionnent, de même que leurs relations telles qu'elles sont exprimées par ces textes. Il est ainsi possible d'automatiser pour de grandes quantités de textes l'extraction d'information en remplaçant efficacement l'humain dans ce processus de production de données numériques. À titre d'exemple, mentionnons le projet *Impresso*, qui a indexé les contenus de plusieurs journaux quotidiens parus sur une période de deux cents ans et qui met à disposition ces données dans une plateforme d'exploration et de recherche [29]. Ou, dans une dimension encore plus futuriste, le projet *Time Machine* [30], qui vise l'indexation et l'extraction automatisées d'information à partir d'archives entières, comme celles de l'ancienne République de Venise, tout en mobilisant les technologies les plus avancées dans toute la chaîne de traitement.

26 Comme indiqué précédemment, l'information ainsi extraite des sources n'est pas « donnée » mais construite – souvent implicitement et donc de manière non optimale – à partir du modèle d'entraînement de la machine qui correspond à la conceptualisation des chercheurs. Elle se situe donc, dans la pyramide, au niveau des factoïdes, car différents documents pourraient mentionner les mêmes faits. C'est à ce stade qu'intervient la deuxième composante, l'interconnexion des entités, appelée également *record linkage* [31], c'est-à-dire l'établissement d'un lien entre l'objet ou l'événement reconnu dans le texte et une ressource clairement identifiée, de même que le fait que les mentions dans différentes sources se réfèrent au même objet. Sur cette base, les factoïdes seront agrégés et transformés en information factuelle. Ce processus peut être également automatisé grâce à l'apprentissage artificiel.

27 Afin de permettre la réutilisation de l'information extraite des sources, il est indispensable d'expliciter et de réfléchir à la conceptualisation qui lui est sous-jacente, car toute donnée numérique est le reflet d'une ontologie, c'est-à-dire d'une sémantique spécifique, qu'on le veuille ou non : « *The opposite of ontology is not non-ontology but just bad ontology* », écrit Giancarlo Guizzardi (2020). Si l'ontologie est standardisée et partagée dans une communauté disciplinaire, il devient possible de comparer les relations entre objets produites par un projet avec celles des autres systèmes d'information, de les aligner et de les intégrer. La condition de réalisation de cette troisième composante du changement de paradigme est l'adoption de modèles conceptuels de référence robustes et adaptés au domaine de recherche concerné, développés en appliquant les méthodologies sémantiques (Staab et Studer 2009 ; Domingue, Fensel et Hendler 2011). Il n'y a qu'à penser au CIDOC CRM (ISO 21127 ; Doerr 2003) et à sa famille d'extensions [32] ou à DOLCE Ultra Light et aux *Ontology Design Patterns* (Presutti et Gangemi 2016), déjà

utilisés dans le domaine de la conservation du patrimoine. Du point de vue des sciences historiques, il s'agit donc de s'approprier ces méthodologies et de produire des modèles extensibles et sémantiquement cohérents.

28 L'identification des mêmes entités dans différents gisements grâce aux référentiels d'autorités, la formalisation de leurs relations en utilisant des ontologies partagées et leur publication sous forme de LOD permettront de reconstituer des pans entiers de l'activité des sociétés du passé afin de les explorer dans leur globalité et de produire un savoir renouvelé. Cette quatrième composante de la transformation numérique va aboutir à la création d'un graphe géant et distribué de l'information historique, un *giant knowledge graph* disciplinaire, qui intégrera également le patrimoine croissant d'éditions textuelles numériques ainsi que les données issues de systèmes d'information pour les biens culturels (Carriero 2019a ; Carriero 2019b [33]). Cette composante ne concerne pas uniquement le volume d'information interrogeable, inimaginable jusqu'ici, mais encore sa diversité et sa qualité dont la vérification pourra être automatisée. Il en va de même pour le processus d'agrégation de l'information factuelle, en laissant aux historiens la tâche de vérifier la qualité et la pertinence du résultat obtenu par la machine [34].

29 Tout le potentiel du passage de l'information présente dans les textes à celle disponible et interrogeable dans les systèmes d'information sémantiques est illustré par le projet *Biografiasampo* [35], qui a transformé les notices de la biographie nationale finnoise en LOD, permettant ainsi une exploration inédite de l'information présente dans les textes (Tamper *et al.* 2021). Ce projet montre que pour pouvoir profiter pleinement du graphe d'information, une cinquième composante est indispensable : l'appropriation par les chercheurs de multiples méthodologies et logiciels d'analyse, et leur intégration réfléchie dans la production du savoir. Mentionnons le blog du Digital Humanities Lab de l'Institut d'histoire européenne de Leipzig [36] ou la nouvelle revue en ligne *Journal of Digital History* parmi les exemples d'utilisation avancée de l'outillage numérique dans la recherche historique [37]. Ces publications documentent sans équivoque le changement de paradigme en cours.

30 Dans ce nouveau contexte, il semble donc indispensable d'introduire la formation aux méthodologies et aux outils numériques dès le début du cursus universitaire. Les humanités numériques se sont développées depuis quelques décennies comme une transdiscipline originale, mais un peu en marge des cursus disciplinaires. Les nouveaux parcours de master ajoutent certes une surcouche importante en méthodologie numérique aux disciplines traditionnelles, mais ils orientent surtout vers les métiers du soutien à la recherche. Comme l'apprentissage de l'outillage disciplinaire est au cœur du paradigme d'une discipline, il faudrait introduire l'enseignement des méthodologies numériques dès le début de la formation, afin de permettre aux futures générations de doctorants, enseignants et chercheurs d'opérer de l'intérieur la transition vers le nouveau paradigme [38]. On pourra ainsi créer une communauté disciplinaire formée aux nouvelles méthodologies, connaissant les enjeux par expérience directe et capable de défendre la place des sciences historiques dans le champ de la science contemporaine.

31 Par ailleurs, l'ampleur de la révolution technologique n'est pas maîtrisable individuellement et il ne s'agit pas de transformer tous les historiens en informaticiens. Il est donc indispensable de mettre en place une infrastructure de recherche disciplinaire en ligne, modulaire, ouverte, interconnectée et portée collectivement, afin de soutenir le travail avec les données numériques en tant que nouveau support de l'information.

L'infrastructure numérique pour la recherche et l'interopérabilité sémantique des LOD

32 La vision exposée ci-dessus est à l'origine du projet *symogih.org* (Système modulaire de gestion de l'information historique), né en 2008 de la volonté de quelques historiens du Laboratoire de recherche historique Rhône-Alpes (Larhra) à Lyon de mutualiser leurs données afin de permettre leur réutilisation [39]. Entre 2007 et 2010, le projet *SIPPAF* (Système d'information patrons et patronat français) – financé par l'Agence nationale de la recherche (ANR) – a mis en place un système d'information prosopographique consacré au patronat français (XIX^e^-XX^e^ siècles [40]). Les données produites continuent à être enrichies et utilisées par des chercheurs et des étudiants plus de dix ans après la fin du financement. Elles ont été réutilisées notamment dans le cadre du projet *Siprojuris* consacré aux professeurs de droit en France de 1804 à 1950 [41]. Ces deux projets disposent chacun d'un site Web dédié bien que la production des données se fasse dans un environnement virtuel de recherche unique, la plateforme *symogih.org*, ce qui favorise la réutilisation. Aussi, la qualité des données est progressivement améliorée et leur durée de vie prolongée bien au-delà de la fin du financement des projets.

33 Les conditions de réalisation d'un tel environnement collaboratif de production d'information (au sens de la pyramide) peuvent s'articuler en deux volets. Premièrement, comme nous l'avons vu, il est indispensable de distinguer soigneusement entre une production d'information visant la factualité et la problématique de recherche qui accompagne la collecte des données. Deuxièmement, il est indispensable de disposer d'une conceptualisation générique et ouverte, dont la publication permet d'expliciter le sens des données et de permettre leur réutilisation. Pour revenir à l'exemple de l'histoire de l'astronomie, l'activité d'enseignement a été modélisée de manière suffisamment générique pour pouvoir être utilisée dans d'autres projets [42].

34 Dans le nouveau contexte du Web sémantique, un processus de réécriture du modèle du projet *symogih.org* a été entamé en 2013, afin de le transformer en ontologie et de publier les données sous forme de LOD (Beretta 2017). Pour l'alignement des entités, nous avons choisi IdRef, le référentiel de l'ABES, comme point d'accès principal au Web sémantique, sans exclure d'autres systèmes de notices d'autorité, ni *Wikidata*, qui est de plus en plus utilisé à cette fin. Une épreuve de concept a été menée avec les données du projet *Siprojuris*, désormais liées au référentiel IdRef et interrogeables sur le point d'accès SPARQL du projet *symogih.org*. Non seulement la liste des publications de chaque professeur de droit est établie en récupérant en temps réel les données des notices du catalogue Sudoc [43], mais encore les données du projet peuvent être enrichies par

celles d'autres ressources, telles BNF Data ou DBpedia [44]. Ce projet réalise ainsi un prototype de la vision des principes FAIR évoquée ci-dessus, sous forme de *linked open research data* (*R-LOD* [45]) : l'information est publiée sous forme de graphes sémantiques interconnectés.

35 Suite à différents échanges avec les experts, il a semblé plus judicieux, en matière d'interopérabilité, d'inscrire l'expérience de modélisation acquise précédemment dans l'univers du CIDOC CRM et de sa famille d'extensions, afin de faciliter l'intégration des données issues de la recherche historique dans un graphe géant d'information (Doerr et Iorizzo 2008). À cette fin, un écosystème d'extensions du CIDOC CRM a été créé dans le projet *Semantic Data for Humanities and Social Sciences* (*SDHSS*) [46], en suivant une méthodologie par couches d'abstraction, afin de produire une sémantique extensible, qui permet d'assurer la cohérence entre les modèles de recherche de différents projets (Beretta 2021 ; Beretta 2022). Pour soutenir et faciliter ce processus, il est nécessaire de disposer d'une application en ligne permettant la modélisation collaborative et l'alignement d'ontologies. Après évaluation des plateformes disponibles, nous avons décidé de développer un nouveau service, OntoME (*Ontology Management Environment* [47]), qui a été adopté par différents projets [48]. À travers une interface machine (API), les espaces de noms et profils applicatifs créés dans OntoME peuvent être affichés et analysés dans Protégé, éditeur d'ontologies développé par l'université de Stanford [49]. On peut donc utiliser ces deux outils de manière complémentaire : OntoME comme support à un travail collaboratif de conceptualisation, Protégé comme outil de vérification du formalisme grâce aux moteurs de *reasoning*.

36 Le consortium *Data for History* [50] a été créé en novembre 2017 lors d'un workshop organisé à l'École normale supérieure de Lyon afin de promouvoir cette vision. Il a été suivi par un deuxième atelier lyonnais en 2018, puis par une rencontre à Leipzig en 2019 [51] et un workshop d'experts en ontologies à Bologne en 2022 [52]. La première conférence internationale (en ligne) en mai-juin 2021 a été organisée par la chaire d'histoire numérique de l'université Humboldt de Berlin [53] et elle a donné lieu à la mise en place d'une série de *Data for History Lectures* [54]. Dans ce contexte s'inscrit aussi un nouveau projet de récupération et sémantisation de données existantes en archéologie et histoire, intitulé *Semantic Data for Humanities and Social Sciences* [55], qui s'appuie, entre autres, sur la collaboration entre l'ABES et le Larhra, notamment dans le projet ANR *HisArc-RDF* « *Partage et réutilisation de données archéologiques et historiques : une description en RDF appuyée sur les référentiels et les normes du Web sémantique* » (2019-2022).

37 En raison de réductions budgétaires et d'autres circonstances, le développement du projet *symogih.org* a été arrêté. Une migration des données est en cours vers *Geovistory* [56], un nouvel environnement virtuel de recherche développé à l'origine par la société KleioLab [57] et désormais porté par le département d'humanités numériques de l'université de Berne [58] en collaboration avec le Larhra dans le cadre d'un partenariat destiné à pérenniser l'infrastructure en constituant un consortium d'institutions publiques. *Geovistory* reprend et développe de manière significative l'approche du projet *symogih.org* tout en apportant des améliorations substantielles. D'une part, la nouvelle application a été réalisée selon la méthodologie UX afin de faciliter la saisie des données : chaque projet

travaille avec une vue sur l'ensemble des données qui lui est propre mais contribue en même temps à enrichir le graphe d'information partagé. D'autre part, le modèle de données est directement géré sur la plateforme OntoME et profite de l'écosystème d'ontologies du projet *SDHSS* en utilisant des profils applicatifs conçus en fonction des besoins de la recherche : dès leur production, les données sont ainsi définies à l'aide d'une sémantique qui garantit leur interopérabilité.

38 Plusieurs modalités d'accès aux données permettent leur exploitation. Une interface de requêtes graphique facilite l'exploration et la préparation des exports en CSV ou JSON. L'analyse des données, réalisée avec les logiciels R ou Python, peut être documentée en utilisant les carnets Jupyter qui, une fois publiés, facilitent la reproductibilité des résultats de la recherche [59]. Les projets peuvent disposer d'un site Web dédié ainsi que d'un point d'accès SPARQL, sur lequel sont publiées les LOD du projet selon sa perspective propre. De plus, un point d'accès SPARQL générique publie les données de l'ensemble de la communauté, avec toute sa richesse en matière de variété de points de vue, voire avec des contradictions possibles [60]. Les entités de *Geovistory* sont reliées, tout comme celles du projet *symogih.org*, avec les notices d'autorité et autres référentiels du Web sémantique et s'inscrivent ainsi pleinement dans la logique des LOD. Cette infrastructure a été mise en place afin de permettre aux projets de recherche en sciences historiques et, plus largement en sciences humaines et sociales, à la fois de gérer une problématique de recherche propre et, en même temps, de participer à la production d'un graphe géant d'information au service de la recherche scientifique et du public.

Conclusion

39 Si elle n'est pas unique, ni une panacée, l'expérience présentée dans ces pages peut être considérée comme une contribution importante à la réalisation de la transition numérique pour les sciences historiques. Mais est-ce qu'un changement de paradigme est réellement en cours ? Il est indéniable que le fait de transférer l'information d'un support analogique vers des infrastructures numériques contenant de plus en plus de données interconnectées et sémantiquement interopérables, par conséquent actionnables pour effectuer des requêtes et analyses de plus en plus sophistiquées, est susceptible de transformer radicalement la manière de produire le savoir historique.

40 Le volume et la qualité de l'information virtuellement disponible sous forme de LOD, ainsi que des documents numériques qu'elles relient, permettent, d'une part, d'envisager des problématiques de recherche jusqu'ici inédites car dépassant ce qu'une seule personne pouvait imaginer pouvoir réaliser au cours d'une vie de recherche. D'autre part, la méthode même de la connaissance est transformée : la problématisation de la recherche et l'ouverture aux sciences sociales, qui ont caractérisé le XX^{e} siècle et permis de dépasser l'histoire-événement et sa dimension descriptive, peuvent désormais être pleinement réalisées, car il est désormais possible, grâce aux gisements de données numériques ouvertes, liées et sémantisées, et aux logiciels d'analyse de plus en plus sophistiqués, de tester des hypothèses explicatives sur un volume d'information

factuelle capable virtuellement de représenter la vie des sociétés du passé dans leur globalité. Une nouvelle appréhension des causes, des structures et des dynamiques des phénomènes historiques devient ainsi envisageable, de même qu'une capacité, je ne dirai pas de prévoir le futur, ce qui est impossible, mais de reconnaître des tendances et de mettre en évidence des dangers et des déséquilibres qui menacent les sociétés contemporaines.

41 On peut subir cette transformation, en risquant une inévitable marginalisation, ou l'assumer en protagonistes. Pour ce faire, il faudra introduire rapidement et le plus conséquemment possible les méthodologies et les outils numériques dans la formation des nouvelles générations d'étudiants sans pour autant négliger les piliers classiques de la discipline, notamment le discernement critique, qui doivent être repensés dans le nouveau contexte numérique. Parallèlement, il est fondamental de soigner l'édification d'une infrastructure disciplinaire multiprojets, distribuée mais sémantiquement interconnectée, afin de tirer tout le profit possible du *graphe géant de l'information historique*, représentation numérique cumulative des sociétés du passé, que les générations successives pourront réutiliser et développer en ouvrant de nouveaux chantiers et problématiques de recherche. Il faut espérer que la communauté des sciences historiques saura assumer les défis que comporte ce changement de paradigme et se donner les moyens de continuer à exercer sa fonction critique dans les nouvelles sociétés numériques.

Bibliographie

Akoka, Jacky, Isabelle Comyn-Wattiau, Stéphane Lamassé et Cédric du Mouza. 2020. « Contribution of Conceptual Modeling to Enhancing Historians' Intuition – Application to Prosopography ». Dans *Conceptual Modeling*, édité par Gillian Dobbie, Ulrich Frank, Gerti Kappel, Stephen W. Liddle et Heinrich C. Mayr, 164-173. Cham : Springer. https://doi.org/10.1007/978-3-030-62522-1_12.

Bécue-Bertaut, Monique. 2018. *Analyse textuelle avec R*. Rennes : Presses universitaires de Rennes.

Beretta, Francesco. 2016. « Pour une annotation sémantique des textes : le projet *symogih.org* et la Text Encoding Initiative ». *Bruniana e Campanelliana. Ricerche filosofiche e materiali storico-testuali* XXII (2) : 453-465. https://doi.org/10.19272/201604102005.

Beretta, Francesco. 2017. « L'interopérabilité des données historiques et la question du modèle : l'ontologie du projet SyMoGIH ». Dans *Enjeux numériques pour les médiations scientifiques et culturelles du passé*, édité par Brigitte Juanals et Jean-Luc Minel, 87-217. Nanterre : Presses universitaires de Paris-Nanterre. https://halshs.archives-ouvertes.fr/halshs-01559816/document.

Beretta, Francesco. 2021. « A Challenge for Historical Research : Making Data FAIR Using a Collaborative Ontology Management Environment (OntoME) ». *Semantic Web* 12 (2) : 279-294. https://doi.org/10.3233/SW-200416.

Beretta, Francesco. 2022. « Interopérabilité des données de la recherche et ontologies fondationnelles : un écosystème d'extensions du CIDOC CRM pour les sciences humaines et sociales ». Dans *Actes des journées « Humanités numériques et Web sémantique » (Nancy, France)*, édité par Nicolas Lasolle, Olivier Bruneau et Jean Lieber, 2-22. https://doi.org/10.5281/zenodo.7014341.

Carriero, Valentina Anita, Aldo Gangemi, Maria Letizia Mancinelli, Andrea Giovanni Nuzzolese, Valentina Presutti et Chiara Veninata. 2019a. « Pattern-Based Design Applied to Cultural Heritage Knowledge Graphs ». *ArXiv:1911.07585 [Cs]*, November. http://arxiv.org/abs/1911.07585.

Carriero, Valentina Anita, Aldo Gangemi, Maria Letizia Mancinelli, Ludovica Marinucci, Andrea Giovanni Nuzzolese, Valentina Presutti et Chiara Veninata. 2019b. « ArCo : The Italian Cultural Heritage Knowledge Graph ». *ArXiv:1905.02840 [Cs]* 11779 : 36-52. https://doi.org/10.1007/978-3-030-30796-7_3.

Cellier, Jacques et Martine Cocaud. 2012. *Le Traitement des données en histoire et sciences sociales. Méthodes et outils*. Rennes : Presses universitaires de Rennes. http://jacquescellier.fr/histoire/site_tdh2/.

Doerr, Martin. 2003. « The CIDOC Conceptual Reference Module : An Ontological Approach to Semantic Interoperability of Metadata ». *AI Magazine* 24 (3) : 75-92. https://doi.org/10.1609/aimag.v24i3.1720.

Doerr, Martin et Dolores Iorizzo. 2008. « The Dream of a Global Knowledge Network – A New Approach ». *Journal on Computing and Cultural Heritage* 1 (1) : 1-23. https://doi.org/10.1145/1367080.1367085.

Domingue, John, Dieter Fensel et James A. Hendler, éd. 2011. *Handbook of Semantic Web Technologies*. Berlin : Springer.

Genet, Jean-Philippe, éd. 2011. *Les Historiens et l'informatique : un métier à réinventer*. Rome : Publications de l'École française de Rome.

Grosjean, Roger. 1972. « Les alignements de Pagliaiu (Sartène, Corse) ». *Bulletin de la Société préhistorique française* 69 (2) : 607-617. https://doi.org/10.3406/bspf.1972.8189.

Guarino, Nicola et Christopher A. Welty. 2009. « An Overview of OntoClean ». Dans *Handbook on Ontologies*, édité par Steffen Staab et Rudi Studer, 201-220. Berlin : Springer.

Guizzardi, Giancarlo. 2020. « Ontology, Ontologies and the "I" of FAIR ». *Data Intelligence* 2 (1-2) : 181-191. https://doi.org/10.1162/dint_a_00040.

Kuhn, Thomas S. 1996 [1962]. *La Structure des révolutions scientifiques* (trad. de l'anglais, éd. augm. de 1970). Paris : Flammarion.

Marrou, Henri-Irénée. 1961. « Comment comprendre le métier d'historien ». Dans *L'Histoire et ses méthodes*, édité par Charles Samaran, 1465-1540. Paris : Gallimard.

Meroño-Peñuela, Albert, Ashkan Ashkpour, Marieke van Erp, Kees Mandemakers, Leen Breure, Andrea Scharnhorst, Stefan Schlobach et Frank van Harmelen. 2015. « Semantic Technologies for Historical Research : A Survey. » *Semantic Web* 6 (6) : 539-564. https://doi.org/10.3233/SW-140158.

Pasin, Michele et John Bradley. 2015. « Factoid-Based Prosopography and Computer Ontologies : Towards an Integrated Approach ». *Literary and Linguistic Computing* 30 (1) : 86-97. https://doi.org/10.1093/llc/fqt037.

Presutti, Valentina et Aldo Gangemi. 2016. « Dolce+D&S Ultralite and Its Main Ontology Design Patterns ». In *Ontology Engineering with Ontology Design Patterns. Foundations and Applications*, édité par Pascal Hitzler, Aldo Gangemi, Krzysztof Janowicz, Adila Krisnadhi et Valentina Presutti, 81-103. IOS Press. https://doi.org/10.3233/978-1-61499-676-7-81.

Puren, Marie et Pierre Vernus. 2022. « Vers une ontologie de domaine pour l'analyse des tissus anciens. Le projet *Silknow* et le cas du patrimoine soyeux européen ». *Humanités numériques* 6. https://doi.org/10.4000/revuehn.3179.

Rowley, Jennifer. 2007. « The Wisdom Hierarchy : Representations of the DIKW Hierarchy ». *Journal of Information Science* 33 (2) : 163-180. https://doi.org/10.1177/0165551506070706.

Schultz, Émilien et Matthias Bussonnier. 2021. *Python pour les SHS. Introduction à la programmation pour le traitement de données*. Rennes : Presses universitaires de Rennes. http://www.pur-editions.fr/detail.php?idOuv=5092.

Staab, Steffen et Rudi Studer, éd. 2009. *Handbook on Ontologies*. Berlin : Springer.

Tamper, Minna, Petri Leskinen, Eero Hyvönen, Fisto Valjus et Kirsi Keravuori. 2021. « Analyzing Biography Collections Historiographically as Linked Data : Case National Biography of Finland ». *Semantic Web Journal*. http://www.semantic-web-journal.net/content/analyzing-biography-collections-historiographically-linked-data-case-national-biography.

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, *et al.* 2016. « The FAIR Guiding Principles for Scientific Data Management and Stewardship ». *Scientific Data* 3 (1) : 160018. https://doi.org/10.1038/sdata.2016.18.

Notes

1 https://fr.wikipedia.org/wiki/Linked_open_data.

2 https://en.wikipedia.org/wiki/Knowledge_graph ; https://en.wikipedia.org/wiki/Google_Knowledge_Graph ; https://fr.wikipedia.org/wiki/Graphe_de_connaissances.

3 https://www.europeana.eu.

4 https://data.bnf.fr.

5 Virtual International Authority File : https://viaf.org.

6 https://scienceplus.abes.fr.

7 https://data.idref.fr.

8 https://isidore.science/sparql ; https://www.nakala.fr.

9 https://www.opendatafrance.net.

10 https://www.science-ouverte.cnrs.fr/dates-cles-science-ouverte/.

11 https://www.ccsd.cnrs.fr/principes-fair/ ; https://www.force11.org/group/fairgroup/fairprinciples/.

12 « *There is an urgent need to improve the infrastructure supporting the reuse of scholarly data* » (Wilkinson *et al.* 2016, Abstract).

13 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.

14 https://en.wikipedia.org/wiki/Upper_ontology.

15 Cet article est issu d'une communication aux *Premières journées historiographiques corses* en juillet 2021.

16 https://www.data.corsica/explore/dataset/liste-des-monuments-historiques-en-corse/table/.

17 https://www.pop.culture.gouv.fr/notice/merimee/PA00099118/.

18 https://fr.wikipedia.org/wiki/SPARQL.

19 Exécuter cette requête sur le serveur SPARQL de *Wikidata*, sélectionner « Graph » dans le menu des visualisations possibles, en dessous de la réponse à gauche, pour l'affichage du résultat sous forme de graphe, ou « Map » pour l'affichage des localités sur la carte de la Corse.

20 https://fr.wikipedia.org/wiki/Palaghju.

21 http://fr.dbpedia.org/page/Palaghju.

22 http://fr.dbpedia.org/page/Roger_Grosjean.

23 https://fr.wikipedia.org/wiki/Roger_Grosjean.

24 https://www.idref.fr/068999836.

25 https://programminghistorian.org/fr/.

26 https://en.wikipedia.org/wiki/Data_curation.

27 Pour le domaine des sciences historiques, voir, par exemple, la *Historical Context Ontology* (*HiCO*), basée sur l'ontologie de documentation de l'origine des données numériques PROV-O : https://marilenadaquino.github.io/hico/. Pour les technologies, voir la nouvelle spécification RDF-star qui est de plus en plus implémentée dans les bases de données sémantiques : https://w3c.github.io/rdf-star/cg-spec/2021-12-17.html.

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

29 https://impresso.github.io.

30 https://www.timemachine.eu.

31 https://en.wikipedia.org/wiki/Record_linkage.

32 https://cidoc-crm.org/collaborations/.

33 https://dati.beniculturali.it/arco-rete-ontologie/.

34 À titre d'exemple, voir Akoka *et al.* (2020).

35 https://biografiasampo.fi (traduire dans la langue de préférence avec Google Translator).

36 https://dhlab.hypotheses.org/2271/.

37 https://journalofdigitalhistory.org/en/issue/jdh001/.

38 Mentionnons à titre d'exemple les travaux pionniers de Cellier et Cocaud (2012), et, plus récemment, Schultz et Bussonnier (2021) ou Bécue-Bertaut (2018).

39 http://symogih.org.

40 http://www.patronsdefrance.fr.

41 http://siprojuris.symogih.org.

42 http://symogih.org/?q=type-of-information-record/97/.

43 http://siprojuris.symogih.org/siprojuris/enseignant/44315/ (onglet « Bibliographie externe »).

44 https://wiki-arhn.larhra.fr/doku.php?id=siprojuris:defi_donnees_2018/.

45 Cf. https://www.switch.ch/connectome/ et https://wiki.esipfed.org/Linked_Open_Research_Data_for_Earth_and_Space_Science_Informatics.

46 https://ontome.net/namespace/11.

47 https://ontome.net.

48 En particulier, deux projets financés sur fonds européens ont utilisé OntoME pour la préparation du modèle de données : *Silknow* (projet ERC, cf. Puren et Vernus [2022]) et *Read-it* (projet JPICH).

49 https://protege.stanford.edu.

50 http://dataforhistory.org.

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

52 https://data4history-unibo.github.io/meeting2022/.

53 https://d4h2020.sciencesconf.org.

54 https://dhistory.hypotheses.org/category/forschungskolloquium/data-for-history-lectures/.

55 http://dataforhumanities.org.

56 https://www.geovistory.org.

57 https://kleiolab.ch.

58 https://www.dh.unibe.ch/index_eng.html.

59 https://github.com/geovistory/HISB-volkzahlung/blob/main/analysis/naissance_origines.ipynb.

60 https://www.geovistory.org/sparql/.

Auteur

Francesco Beretta

UMR 5190 Laboratoire de recherche historique Rhône-Alpes (Larhra), CNRS, Lyon, France
Francesco Beretta est chargé de recherche au CNRS et d'enseignement en méthodologies numériques pour les sciences historiques à l'université de Neuchâtel. Spécialiste d'histoire des sciences à l'époque moderne, il est actif depuis plus de quinze ans dans le domaine des systèmes d'information pour la recherche, les ontologies et l'interopérabilité des données de la recherche (projets *symogih.org*, *ontome.net*, *dataforhistory.org*, *geovistory.org*).
ORCID 0000-0002-4389-4126
francesco.beretta@cnrs.fr

Droits d’auteur