Sens public

Qu'y a-t-il dans un nom ?
Les données ouvertes liées comme base d'une écologie pour la publication scientifique, dynamique et décentrée

John Simpson et Susan Brown

2021

URI : https://id.erudit.org/iderudit/1089588ar
DOI : https://doi.org/10.7202/1089588ar

Aller au sommaire du numéro

Éditeur(s)

Département des littératures de langue française

ISSN

2104-3272 (numérique)

Découvrir la revue

Citer cet article

Simpson, J. & Brown, S. (2021). Qu'y a-t-il dans un nom ? Les données ouvertes liées comme base d'une écologie pour la publication scientifique, dynamique et décentrée. *Sens public*, 1–21. https://doi.org/10.7202/1089588ar

Résumé de l'article

Les données ouvertes liées peuvent produire un environnement scientifique plus interconnecté et plus navigable qui permet : une meilleure intégration des matériaux de recherche ; la possibilité d'aborder les spécificités de la nomenclature, des discours et des méthodologies ; et la capacité de respecter les investissements institutionnels et individuels. Cet article propose une écologie de la publication des données liées basée sur la mise en place de collaborations entre les communautés des chercheurs, des éditeurs et des bibliothèques. Cette vision est tempérée par l'état des pratiques de publication des données liées et les lacunes infrastructurelles en ce qui concerne la possibilité de telles collaborations, en particulier dans le domaine des sciences humaines et sociales. **Addendum :** Cet article a été publié à l'origine dans Scholarly and Research Communication en 2015. Bien que certaines des informations ne soient plus aussi pertinentes, l'article présente une problématique qui a depuis été abordée par un certain nombre d'initiatives, notamment les projects Linked Infrastrucutre for Networked Cultural Scholarship (LINCS), Revue 2.0, et Information Economy Meta-Language (IEML).

Sens [public]

Qu'y a-t-il dans un nom ?

Les données ouvertes liées comme base d'une écologie pour la publication scientifique, dynamique et décentrée

John Simpson, Susan Brown

Publié le 01-03-2021

http://sens-public.org/articles/1484

Résumé

Les données ouvertes liées peuvent produire un environnement scientifique plus interconnecté et plus navigable qui permet : une meilleure intégration des matériaux de recherche ; la possibilité d'aborder les spécificités de la nomenclature, des discours et des méthodologies ; et la capacité de respecter les investissements institutionnels et individuels. Cet article propose une écologie de la publication des données liées basée sur la mise en place de collaborations entre les communautés des chercheurs, des éditeurs et des bibliothèques. Cette vision est tempérée par l'état des pratiques de publication des données liées et les lacunes infrastructurelles en ce qui concerne la possibilité de telles collaborations, en particulier dans le domaine des sciences humaines et sociales.

Addendum : Cet article a été publié à l'origine dans Scholarly and Research Communication en 2015. Bien que certaines des informations ne soient plus aussi pertinentes, l'article présente une problématique qui a depuis été abordée par un certain nombre d'initiatives, notamment les projects Linked Infrastrucutre for Networked Cultural Scholarship (LINCS), Revue 2.0, et Information Economy Meta-Language (IEML).

Abstract

Linked open data can produce a more interconnected and navigable scientific environment that enables: better integration of research materials; the ability to address the specifics of nomenclature, discourse, and methodologies; and the ability to respect institutional and individual investments. This paper proposes an ecology of linked data publication based on the establishment of collaborations between the communities of researchers, publishers, and libraries. This view is tempered by the state of practice in publishing linked data and infrastructural gaps in the potential for such collaborations, particularly in the humanities and social sciences.

Addendum: This article was originally published in Scholarly and Research Communication in 2015. While some of the information therein may not be as pertinent today, it presents a challenge that has since been undertaken by a number of initiatives, most notably the Linked Infrastructure for Networked Cultural Scholarship (LINCS) project, Revue 2.0, and the Information Economy Meta-Language (IEML).

Mot-clés : connaissance du public, diffusion des connaissances, production du savoir, humanités numériques, gestion de contenu

Keywords: public knowledge, knowledge dissemination, knowledge production, digital humanities, content management

Table des matières

Qu'y a-t-il dans un nom ?

John Simpson Susan Brown

Introduction

Traduit de l'anglais par Jasmine Drudge-Willson (version originale CC BY-NC-ND)

L'isolement vis-à-vis des documents connexes nuit aux ressources scientifiques diffusées en ligne, à l'intérieur et à l'extérieur des sciences humaines. Une des principales plaintes des chercheurs en ce qui concerne la recherche et l'utilisation des documents numériques est qu'ils existent isolément, et qu'ils ne sont pas reliés aux autres documents pertinents (Bulger et al. 2011). Il en va de même pour les publications liées à l'impression conventionnelle – comme les revues en ligne, les livres numérisés et les livres électroniques – dont l'accès est souvent exacerbé par des murs de paiement ou par les structures de base de données dans lesquelles elles sont hébergées : les projets des humanités numériques publiés sur le Web par des particuliers ou des bibliothèques, ainsi que les grands projets de numérisation ou d'agrégation. Les services d'indexation contribuent à atténuer ce problème, mais l'interconnexion des ressources avec les documents qu'elles citent et les documents qui les citent demeure un défi. Il en résulterait des avantages énormes si nous pouvions, par exemple, tirer parti des réseaux de citation et les formaliser dans notre environnement d'information, et ce, que les ressources dans lesquelles les citations se trouvent soient publiées officiellement ou non, qu'elles se trouvent dans les commentaires des utilisateurs ou dans les discours qui les accompagnent sur les réseaux sociaux. À la base, utiliser le discours des entités ou des « objets » associés au Web sémantique ou aux données ouvertes liées (DOL, ou LOD en anglais) implique la capacité d'interconnecter les diverses entités liées à

ces ressources entre elles (Learn More about « Learn More About WorldCat Works » 2015).

La quantité et la diversité des discours scientifiques largement déconnectés qui circulent sous forme numérique constituent un défi et une occasion sans précédent. Relever ce défi de manière réalisable permettrait de faire deux choses importantes. Premièrement, un niveau plus élevé d'interconnexion et d'interopérabilité des textes et des contextes contribuerait grandement à résoudre « le problème d'un million de livres » de Gregory Cran (2006) [1]. Cela permettrait à la recherche scientifique de se développer d'une manière qui, jusqu'à présent, n'a été accessible qu'à une très petite minorité de chercheurs en sciences humaines disposant du financement et des compétences nécessaires pour rassembler de vastes ensembles de données pour leur propre usage. Même ces efforts ont été inévitablement limités par le fait que leurs ensembles de données, bien que larges, demeuraient néanmoins limités. Deuxièmement, le discours scientifique interconnecté et imbriqué a de bonnes possibilités d'accroître son impact, mais ce travail est actuellement invisible dans les principaux moteurs de recherche et il ne trouve pas sa place parmi les autres sources d'information qui remplissent le Web. Il s'agit là d'un sujet de regret, étant donné sa pertinence dans de nombreux débats contemporains et sa plus grande exigence d'autorité et de fiabilité que nombre de ses sources actuelles.

Cet article aborde le problème plus modeste et plus gérable de l'*interconnexion* comme un premier pas crucial vers l'interopérabilité en proposant les DOL, incluant l'exploitation des entités et des relations, comme moyen de produire un environnement de connaissances plus interconnecté et plus facilement navigable. Les éléments de base nécessaires à la mise en place d'un tel système existent, et des initiatives clés sont d'ailleurs en cours au sein des bibliothèques, des musées et du milieu de l'édition. L'accent sera mis ici sur la communauté scientifique et sa capacité à s'engager dans ces développements de manière à renforcer la forme générale du Web sémantique et à aider les humanités numériques à surmonter certains obstacles majeurs qui ont entravé leur impact tant dans les humanités traditionnelles que dans l'environnement plus large de l'information. Nous n'employons pas une métaphore environnementale – l'écologie – au mépris des effets extrêmement néfastes des déchets électroniques et de la consommation d'énergie à l'échelle mondiale (« Digital Environmental Humanities » 2015

1. "The million books problem"

; Uddin et Rahman 2011 ; Widmer et al. 2005), ni pour « brouiller » les caractéristiques et effets locaux de ce que nous décrivons (Jaeger et al. 2009). La métaphore de l'écologie de publication met en évidence plusieurs aspects de cette approche.

Selon la définition initiale du disciple darwinien Ernst Haeckel, l'écologie considère « les relations de l'organisme avec l'environnement, y compris, au sens large, toutes les conditions d'existence » [2] (cité dans Ergerton (2013, 226)). L'application d'un cadre écologique souligne à quel point toute tentative de modification des communications et des discours scientifiques doit être comprise en termes de diversité et de systématisation, car elle implique de modifier les liens entre les personnes et les conditions matérielles et institutionnelles dans lesquelles elles travaillent. Comme Bonnie A. Nardi et Vicki O'Day (1999) ont argumenté en introduisant le terme :

> Une écologie d'information est un *système* complexe de parties et de relations. Elle présente de la *diversité* et connaît une évolution continue. Les différentes parties d'une écologie changent ensemble selon les relations dans le système au cours d'une *co-évolution*. Plusieurs *espèces clés* nécessaires à la survie de l'écologie sont présentes. Les écologies d'information ont une notion de *localité* (n.p.). [3]

Cadrer ce problème comme un problème écologique nous permet également de penser en termes d'« écotones », « une région d'interface entre deux écosystèmes différents » [4] (Hegde 2012) – c'est-à-dire de régions dynamiques où le mélange des populations en marge de deux communautés différentes produit des pressions inhabituelles et stimule le changement. Le présent article identifie certaines des caractéristiques des écotones associées aux zones bordières entre les communautés de l'édition scientifique et des bibliothèques ; les secteurs de l'érudition citoyenne, des archives, des galeries et des musées

2. "The relations of the organism to the environment including, in the broad sens, all the"conditions of existence""

3. "An information ecology is a complex *system* of parts and relationships. It exhibits *diversity* and experiences continual evolution. Different parts of an ecology *coevolve*, changing together according to the relationships in the system. Several *keystone species* necessary to the survival of the ecology are present. Information ecologies have a sense of *locality*" (n.p.).

4. "An interface region between two different ecosystems"

mériteraient d'être examinés de façon similaire. Les écotones sont considérés comme essentiels pour soutenir « diverses communautés et… [pour affecter] la circulation des matériaux qui traversent le terrain » [5] (Risser 1990, 9), qui résonnent avec les préoccupations entourant le Web sémantique émergeant (Brown et Simpson 2013). Les espaces bordiers ne sont pas des espaces vides, mais des zones fertiles, voire conflictuelles, qui sont cruciales pour le développement d'un environnement d'information solide et équilibré (Brown 2011). Toutes aussi pertinentes sont les connotations de l'écologie en tant que mouvement social, le sentiment qu'il existe de meilleures et de pires façons d'influer sur un environnement, et que les interventions devraient être bénéfiques dans leurs conséquences à long terme au-delà du contexte immédiat.

Les avantages d'une écologie de connaissances des données ouvertes et liées

Alors, comment les données liées peuvent-elles conduire à une meilleure écologie de la publication pour l'érudition et, en particulier, permettre aux publications scientifiques d'interagir avec les ensembles de données produits par les bibliothèques et les musées, d'une part, et les entreprises de publication officielle, d'autre part, afin de les améliorer et de les enrichir ? Ici, l'accent sera mis sur plusieurs avantages qui n'épuisent aucune possibilité : 1) l'interconnexion et, au moins au niveau de l'interface, l'intégration des ressources ; 2) la mise à disposition d'informations contextuelles et relationnelles comme base d'un riche environnement de connaissances ; 3) les boucles de rétroaction qui améliorent la qualité des données, en particulier celles qui sont fournies par les fournisseurs d'information à grande échelle ; et 4) l'incorporation d'une diversité de discours, de méthodologies et de données, y compris des ontologies nuancées et des ensembles de données qui respectent le local et le particulier avec les valeurs aberrantes qui peuvent apparaître comme du « bruit » dans de grands ensembles.

5. "Diverse communities and… [affecting] the flow of materials across the landscape"

L'interconnexion et, au moins au niveau de l'interface, l'intégration des ressources

Il s'agit du cas d'utilisation prééminent ou général pour les applications des DOL dans les domaines liés aux sciences humaines. Comme le soutient Jim Hendler (2011), le cadre de description des ressources (RDF) du Web sémantique a bien compris ce que le langage XML (*Extensible Markup Language*) a mal compris : les liens externes. Beaucoup d'énergie se concentre actuellement sur le potentiel des DOL pour aider l'exposition et l'intégration de grands ensembles de données. Les bibliothèques et les musées sont les secteurs où ce type d'initiatives est le plus important avec, notamment, les projets pilotes d'Europeana LOD (2014) et la British Museum Collection of RDF (« British Museum Collection » 2015). Plus près de nous se trouve la preuve de concept « Au-delà des tranchées » élaborée par le Réseau pancanadien du patrimoine documentaire (Price 2012), y compris les principales bibliothèques de recherche et Canadiana (Wuppleman 2012) et, plus récemment, le novateur Munnin Project qui utilise des DOL pour simuler des tranchées de la Première Guerre mondiale (« The Muninn Project » 2015 ; Warren 2012). Aux États-Unis, l'initiative Linked Data for Libraries (LD4L) et le projet VIVO utilisent également des données liées pour agréger des données scientifiques et des collections des bibliothèques, en tirant parti des ressources bibliothécaires ouvertes tel le Virtual International Authority File (« VIVO Open Research Networking Community Group » 2015 ; « Linked Data for Libraries (LD4L) » 2014).

Tous ces projets représentent des cas d'utilisation convaincants pour l'utilisation des données liées afin d'exposer et d'interconnecter les résultats de recherche et les réseaux de publication des chercheurs. Aucun n'intègre toutefois l'activité de recherche dans la vision de l'écologie de publication qui en résulte. L'Online Computer Library Center (OCLC) a réalisé un certain travail en collaboration avec des chercheurs (Godby et al. 2019), mais a reconnu les obstacles qui entravent ce type de collaborations. Apparemment, sa collaboration avec la communauté Wikipédia serait mieux établie et même automatisée (Klein 2012 ; Smith-Yoshimura, Michelson, et Mardutho 2013 ; OCLC Research 2014). S'il existe certainement quelques exceptions (par exemple le projet Digitised Manuscripts to Europeana (DM2E), qui est lié à l'initiative infrastructurelle Digital Research Infrastructure for the Arts and Humanities (DARIAH-EU)), les projets de recherche actifs sont généralement

omis des processus et des flux de travail visant la production et la publication de vastes ensembles de données sur les objets des sciences humaines.

L'inclusion de chercheurs et de projets de recherche actifs demandent de reconsidérer nos manières de concevoir la stabilité des sources et les limites des archives ; pourtant, les omettre représente une occasion manquée pour enrichir ces ressources.

La mise à disposition d'informations contextuelles et relationnelles comme base d'un riche environnement de connaissances

Étant donné les attentes élevées à l'égard de l'actualité des ressources du Web, l'interconnexion des documents de recherche scientifique avec les ensembles de données publiés fournirait des renseignements contextuels utiles pour ces ensembles de données, puisque les travaux scientifiques mettent en relation les sources primaires et les travaux publiés avec les débats d'actualité. Comme l'a dit l'informaticien R. J. Searle, les humanistes, dans un sens, « sont les conservateurs par excellence de l'information savante » [6] parce qu'ils transforment les données primaires « brutes » en contenu « institutionnel » secondaire (cité dans Benardou et al. (2010, 28)). Il y a beaucoup à gagner d'une meilleure intégration des matériaux de recherche avec les sources primaires et secondaires sur lesquelles ils s'appuient. Au-delà de l'interconnexion avec des ressources externes pour l'information contextuelle, des normes émergentes comme l'Open Annotation Data Model (« Open Annotation Data Model » 2013) offrent la possibilité pour les éditions en ligne de textes littéraires primaires, par exemple, de s'appuyer sur des notes de recherche produites par des chercheurs dans d'autres contextes.

Des boucles de rétroaction qui améliorent la qualité des données, en particulier celles qui émanent des fournisseurs d'information à grande échelle

Les chercheurs ont l'expertise et la motivation nécessaires pour corriger les données douteuses qui circulent. Certains projets novateurs, comme l'Early Modern OCR Projet (eMOP), jettent des ponts entre les fournisseurs de contenu numérique à grande échelle et la communauté des chercheurs, et ce, pour leur bénéfice mutuel. Ces efforts peuvent inciter les chercheurs à corriger les erreurs en vue d'améliorer les efforts de numérisation à grande

6. "Are curators par excellence of scholarly information"

échelle en permettant aux utilisateurs de corriger la reconnaissance optique de caractères (ROC, OCR en anglais), ou de noter les images numérisées de mauvaise qualité intégrées dans les collections. Ce qu'il faut, ce sont des outils qui permettent aux fournisseurs de données de recueillir facilement des données rétrospectives sur les corrections dans leurs ensembles de données sources, d'agréger ces informations au sein d'interfaces en fonction de leur provenance et de critères de confiance et d'intégrer les résultats par apprentissage automatique dans les processus de ROC afin d'améliorer l'exactitude globale.

L'incorporation d'une diversité de discours, de méthodologies et de données

Les sciences humaines ont beaucoup à apporter au développement d'une écologie de données liées plus large dans le domaine des ontologies nuancées et des ensembles de données qui respectent le local et le particulier, y compris les valeurs aberrantes qui peuvent apparaître comme du « bruit » dans de grands ensembles de données. La possibilité d'aborder les spécificités de la nomenclature, des discours et des méthodologies des disciplines et des sous-disciplines des sciences humaines tout en les reliant – et la capacité de respecter les investissements institutionnels et individuels dans la propriété ou le crédit des ressources en permettant la collecte de données identifiables tout en favorisant l'interconnexion des ressources – contrecarreront la tendance des DOL à masquer la différence et la diversité résultant du processus de mise à l'échelle.

La modélisation d'une écologie ouverte

Comme point de départ, nous proposons un modèle de très haut niveau pour une écologie de publication décentrée et dynamique basée sur la mise en place de collaborations entre les communautés des chercheurs, des éditeurs et des bibliothèques fondées sur des principes de données liées (se référer à la figure 1).

Les lignes de couleur pleines entre les catégories de contenu représentent le degré élevé de complémentarité des données détenues et la capacité de chaque

domaine d'améliorer les autres de différentes façons. Il s'agit d'une approche plus suggestive que compréhensive. Chaque domaine n'est contenu que de façon minimale dans une forme poreuse semblable à un nuage qui se chevauche avec les autres, et au-dessus d'eux se trouvent les services de données liées qui sont essentiels à une écologie dynamique et productive comme celle que nous avons envisagée. Les flèches vertes pointillées qui se déplacent dans les écotones entre les domaines illustrent à quel point les synergies indiquées par les flèches pleines présupposent de tels services, mais ils ne sont pas encore disponibles.

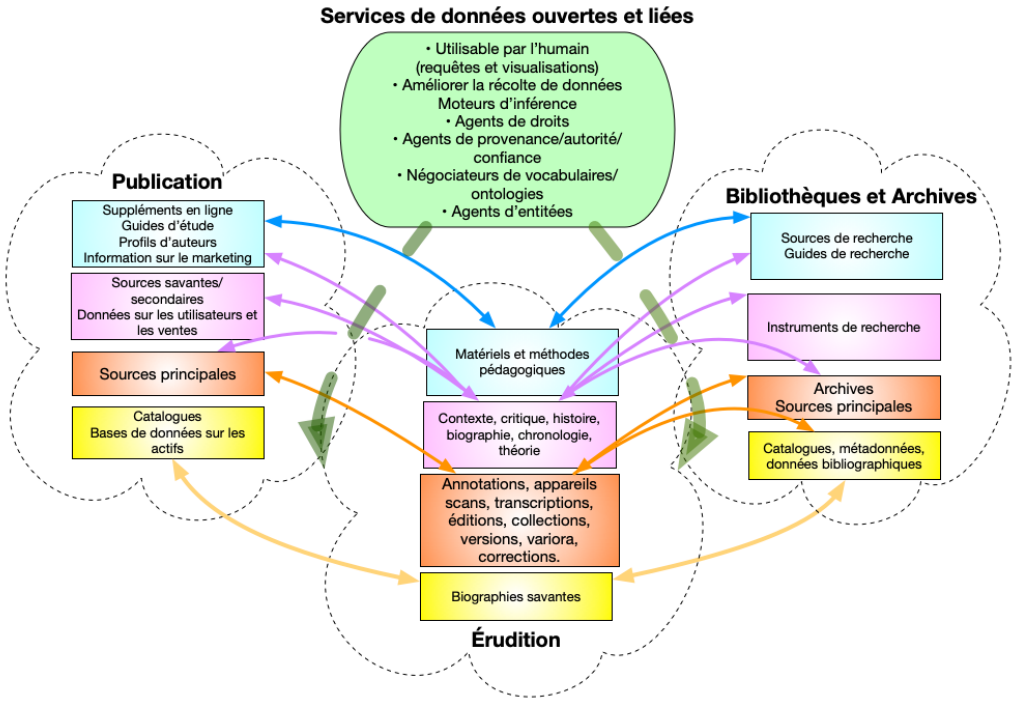

FIGURE 1 – Figure 1 : Esquisse de l'écologie de la publication scientifique dynamique basée sur les données ouvertes et liées

Les lacunes de fonctionnalité

Comme l'indiquent les flèches pointillées, la vision de la gloire que le Web sémantique pourrait offrir doit être tempérée par une prise en compte de l'état actuel des pratiques et de l'infrastructure de publication des données liées.

Il faut combler d'importantes lacunes en ce qui concerne les outils et l'infrastructure avant que ce modèle ne devienne réalité. Nous nous concentrons ici sur deux lacunes complémentaires dans l'écologie de publication des : le raffinement des entités et la nuance des ontologies qui les relient.

Désambiguïsation/alignement/liaison des entités

La conversion ou l'agrégation entièrement automatisée des matériaux existants en DOL produit des résultats qui effacent les distinctions et les différences autour desquelles une grande partie du travail en sciences humaines s'articule. Le refus du traitement automatisé peut expliquer pourquoi les ensembles de données « liées » des sciences humaines sont souvent autoréférentiels, avec peu ou pas de liens avec des données externes. Il existe un besoin urgent de technologies DOL permettant une surveillance humaine efficace, un perfectionnement et une correction des processus automatisés afin de garantir la création ou l'adaptation d'ensembles de données liées en lesquels les humanistes ont confiance. Ce qu'il faut, c'est un flux de travail permettant aux chercheurs de prendre un ensemble de données structuré ou non structuré existant et d'effectuer une série d'opérations pour le préparer en tant que DOL. Les opérations sont les suivantes : 1) effectuer la reconnaissance/extraction des entités nommées et des triples sur l'ensemble de données, ce qui peut impliquer l'utilisation d'ensembles d'entraînement pour obtenir des résultats précis ; 2) faire correspondre les résultats aux collections de DOL existantes qui seront sélectionnables/configurables par l'utilisateur ; 3) présenter aux utilisateurs des correspondances candidates pour les entités et les triples ambigus afin de traiter les correspondances imparfaites et les triples candidats ; 4) à partir de ces données, produire des annotations DOL des données ou intégrer des identificateurs DOL dans les données (ceci est essentiel pour les projets de sciences humaines avec des métadonnées intégrées), en s'appuyant sur l'Open Annotation Data Model (« Open Annotation Data Model » 2013) ; et 5) retransmettre les résultats dans un système d'apprentissage automatique afin d'améliorer les correspondances futures.

Des composants *open source* pour un tel flux de travail existent dans des outils tels le Stanford Named Entity Tagger (The Stanford Natural Language Processing Group s. d.) et le LODE (s. d.) – le Linked Open Data Enhancer développé en partenariat avec l'Indiana Philosophy Ontology (InPho) Pro-

ject (s. d.). Ce qui n'existe pas, c'est un flux de travail utilisable et accessible qui pourrait servir à un large éventail de types de textes. Un tel flux de travail permettrait de faire progresser un certain nombre de projets de recherche existants en matière de DOL. Il comblerait une lacune importante dans l'infrastructure en permettant l'interconnexion des données des éditeurs, des bibliothèques et des musées avec les données scientifiques afin de créer un ensemble de relations riches et symbiotiques. En outre, un tel flux de travail encouragerait l'utilisation des DOL par les humanistes, en poussant les données des sciences humaines à de nouveaux niveaux d'interopérabilité tout en améliorant les ensembles de données existants et en permettant de nouveaux types de recherches et d'inférences dans les ensembles de données culturelles. L'absence d'un tel outil est également ressentie par les principaux fournisseurs d'information. Des organisations qui fournissent les ensembles de données d'autorité ultime dans notre domaine – comme le Library of Congress et l'OCLC, le centre de bibliothèque informatique en ligne à but non lucratif qui héberge WorldCat – seront sollicitées pour la désambiguïsation des entités de données liées, mais la production de données liées reste entravée par l'absence des processus décrits ici. Par exemple, l'OCLC publiera bientôt environ 100 millions de noms de personnes sous la forme de données liées, en plus des noms existants et des 197 millions de titres d'œuvres déjà publiés. Cependant, pour générer cet ensemble de données, l'OCLC a choisi d'ignorer les correspondances imparfaites ; par exemple, les auteurs qui présentent de légères variations dans la représentation de leur nom (par exemple, « E. Pauline Johnson » contre « Pauline Johnson ») ne seront pas considérés comme appartenant à la même entité (Fons 2014). Lors de la réunion de la Coalition for Networked Information à l'automne 2014, les directeurs de grands projets de recherche sur les DOL ont convenu que des services de réconciliation sont nécessaires de toute urgence ; pourtant, personne au sein cette communauté ne s'est engagée à produire un tel outil.

Bien que relativement modeste et tout à fait faisable, un flux de travail utilisable et généralisé de ce type pourrait changer la donne. Comme l'affirme Dominic Lam (2014), de tels flux de travail sont essentiels à l'intensification de la recherche dans le domaine des humanités numériques. De plus, au fur et à mesure que les technologies du Web sémantique deviendront plus répandues (comme dans le cas de Google), l'impact public de l'exposition et de l'interconnexion de grandes quantités de données sur les sciences humaines pourrait être considérable.

Naviguer entre ontologies

Une enquête que nous avons réalisée sur la mise en œuvre des ontologies sur le Web sémantique montre que le graphique de l'utilisation des ontologies a une très longue traîne, ce qui suggère qu'une convergence accrue dans l'adoption des ontologies serait nécessaire si l'objectif était un Web interrelié (Simpson, Brown, et Goddard 2013). La flexibilité de la technologie des données liées réside dans le fait que chaque entrepôt de données peut développer son propre vocabulaire et sa propre ontologie pour répondre à ses besoins, tout en étant relié à d'autres entrepôts de données. Cependant, la connexion avec d'autres données signifie de relier une ontologie à une autre, ce qui entraîne une pression à la généralisation plutôt qu'à la spécificité. Ce n'est pas un hasard si le vocabulaire RDF le plus couramment utilisé est le Dublin Core Metadata Initiative (« Dublin Core Metada Initiative » 2015), dont le succès peut être attribué en grande partie à sa grande simplicité et son applicabilité très large (Simpson, Brown, et Goddard 2013). Pourtant, la généralisation rend les données beaucoup moins utiles pour la recherche en sciences humaines, ce qui implique l'utilisation de « *juke-boxes* d'information » (McCarty 2005) plutôt que d'outils de recherche nuancés. Des initiatives tel le Linked Data for Libraries (« Linked Data for Libraries (LD4L) » 2014) s'associent à des ontologies majeures tels schema.org et Friend of a Friend (FOAF) afin d'assurer une exposition à travers les grands moteurs de recherche. Bien qu'il s'agisse en soi d'un objectif logique et louable, il implique des compromis, y compris le fait de déformer ou de déspécifier certaines des caractéristiques des ontologies développées spécifiquement pour les données bibliographiques afin de les « adapter » à l'ontologie dominante (Krafft et Cramer 2014). Si de telles normes occultent même les catégories relativement simples des principales normes de catalogage, quelle part des caractéristiques éclectiques, nuancées et plus précises des ensembles de données ouvertes liées aux sciences humaines sera perdue lorsqu'il s'agira d'aligner les ontologies ?

Ce qui est nécessaire, c'est un ensemble d'outils pour l'accès aux données liées afin d'aider les chercheurs et les spécialistes de l'information à sélectionner les ensembles de données, à cerner les différences importantes à l'intérieur et entre eux, et à naviguer dans ces différences en fonction des besoins méthodologiques particuliers de leur enquête. L'outil permettrait de faire le lien entre des ensembles de données entiers au choix de l'utilisateur, de contrôler la façon dont les ontologies RDF sont mobilisées et, par la suite, de contrôler

la manière dont les inférences sont faites. Il est essentiel d'établir des liens entre les dépôts de données scientifiques pour qu'ils conservent une partie de la richesse de leurs ontologies locales afin d'éviter une généralisation excessive des ontologies du Web sémantique. Prenons l'exemple d'une chercheuse intéressée par l'exploration de la question complexe et non résolue de l'utilisation des pseudonymes par les femmes écrivains et de leur relation avec l'histoire de la réception. Cette chercheuse pourrait travailler avec les données provenant d'un certain nombre d'ensembles de données de recherche existants sur l'écriture des femmes, qui contiennent tous un riche contenu de réception et des informations très détaillées sur les pseudonymes. L'outil lui permettrait de comparer les ontologies de ces collections à celles d'ensembles de données plus généraux comme le Virtual International Authority File et DBpedia, en notant les différences dans le traitement des noms de personnes. Elle pourrait refuser d'aller vers un dénominateur commun en aplatissant tous les types de noms dans un rôle de « créateur », en choisissant de conserver une plus grande granularité dans les modèles de données associés aux collections de recherche. L'outil permettrait de « resserrer » en s'appuyant sur des relations plus précises pour informer des relations plus générales. Les décisions de la chercheuse seraient éclairées par la possibilité de sélectionner des entités échantillons pour les auteurs qu'elle connaît et d'observer les conséquences de ses choix dans les données de sortie, ce qui lui permettrait de regrouper les matériaux ou de déduire des triples différemment. Les choix de la chercheuse pourraient être sauvegardés dans la bibliothèque de l'outil, pour une utilisation ultérieure par elle ou d'autres personnes, afin de documenter son processus de recherche. Ce type d'engagement de spécialistes dans les ontologies pourrait vraisemblablement aller à l'encontre des tendances homogénéisantes du Web sémantique, si une boucle de rétroaction pouvait être créée pour recueillir les résultats de travaux fiables afin de respecter les relations qui ont été trop généralisées dans la production des données liées, ou d'enrichir des jeux de données plus spécifiques qui n'étaient pas précis ou complets au départ.

Conclusion

Cette discussion n'épuise en rien les lacunes. Le modèle indique une gamme de services DOL nécessaires, dont la plupart n'existent pas encore ou, du moins, n'existent pas sous la forme mature et généralisée nécessaire pour sou-

tenir le type d'échange dynamique de DOL envisagé ici. Il s'agit notamment d'améliorer les mécanismes permettant d'établir des conditions automatisées pour évaluer la provenance, l'autorité et la fiabilité des ressources de DOL, ainsi que des outils pour recueillir et intégrer les corrections et les améliorations. Les droits sont bien sûr une considération majeure. Il reste aussi le fait qu'en dépit de quelques belles interfaces sur mesure adaptées à des collections spécifiques, nous manquons de très bonnes interfaces utilisables par l'humain pour le Web sémantique en général, que ce soit pour des requêtes qui s'appuient sur la structure sémantique ou des visualisations de parties du graphique. Nous soulignons ici deux lacunes que nous considérons particulièrement importantes pour la communauté des sciences humaines. L'élément le plus immédiatement accessible pour le travail dans ce domaine réside dans l'identification et la mise en relation des entités, ce qui permettrait aux données des sciences humaines de passer au Web sémantique et de constituer un élément majeur des sciences humaines en contact avec le public. Un outil de négociation des ontologies, ou ce que nous aimons considérer comme une « machine à différences » (en hommage à Charles Babbage), pourrait être la contribution la plus significative des sciences humaines à l'écologie émergente du Web sémantique, surtout si elle pouvait enrichir les ontologies dans d'autres domaines comme l'édition ou les bibliothèques. Une approche de la publication scientifique numérique fondée sur les entités permettrait d'intégrer l'érudition vivante aux côtés des ressources imprimées, ce qui reflèterait la nature toujours plus dynamique de la production scientifique à l'ère numérique en tant que composante nécessaire de l'environnement du savoir en ligne. Elle offrirait aux chercheurs numériques des solutions locales en matière de contrôle de l'autorité, de recherche d'informations, de visualisation d'informations et, à plus long terme, d'inférence et de raisonnement faisant appel à d'autres sources de connaissances. En bref, elle représenterait une opportunité de collaborations fructueuses avec d'autres secteurs étroitement liés à l'économie du savoir, combinée à la possibilité d'influencer plus directement le Web en tant qu'espace évolutif de production et de diffusion du savoir.

Bibliographie

Benardou, Agiatis, Panos Constantopoulos, Costis Dallas, et Dimitris Gavrilis. 2010. « Understanding the information requirements of arts and humanities scholarship ». *International Journal of Digital Curation* 5 (1) : 18-33.

« British Museum Collection ». 2015. https://old.datahub.io/dataset/british-museum-collection.

Brown, Susan. 2011. « Don't Mind the Gap : Evolving Digital Modes of Scholarly Production across the Digital-Humanities Divide ». In *Retooling the humanities : The culture of research in Canadian universities*, édité par Daniel Coleman et Smaro Kamboureli, 203-31. Edmonton : University of Alberta Press. http://hdl.handle.net/10402/era.25382.

Brown, Susan, et John Simpson. 2013. « The curious identity of Michael Field and its implications for humanities research with the semantic web ». In *2013 IEEE International Conference on Big Data*, 77-85. IEEE. http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6691674&tag=1.

Bulger, M, E Meyer, De la FlorG, M Terras, S Wyatt, M Jirotka, K Eccles, et others. 2011. « Reinventing research ? Information practices in the humanities ». *Information Practices in the Humanities (March 2011). A Research Information Network Report*.

Crane, Gregory. 2006. « What do you do with a million books ? » *D-Lib magazine* 12 (3).

« DBpedia ». 2015. https://wiki.dbpedia.org/.

« Digital Environmental Humanities ». 2015. https://dig-eh.org/.

« Dublin Core Metada Initiative ». 2015. https://www.dublincore.org/.

Egerton, Frank N. 2013. « History of ecological sciences, part 47 : Ernst Haeckel's ecology ». *The Bulletin of the Ecological Society of America* 94 (3) : 222-44.

« eMOP : Early Modern OCR Project ». 2015. https://emop.tamu.edu/.

Europeana. 2014. « Linked Open Data ». *Europeana Pro*. https://pro.europeana.eu/page/linked-open-data.

Fons, Ted. 2014. « Transforming bibliographic records into linked open data (LOD) ». *Panel presentation at the Coalition for Networked Information Fall 2014*. https://www.cni.org/topics/information-access-retrieval/exposing-library-collections-on-the-web-challenges-and-lessons-learned.

Godby, Jean, Karen Smith-Yoshimura, Bruce Washburn, Knudson DavisKalan, Karen Detling, Fernsebner EslaoChristine, Steven Folsom, et al. 2019. « Creating Library Linked Data with Wikibase : Lessons Learned from Project Passage ». OCLC Research Report. https://www.oclc.org/content/dam/research/publications/2019/oclcresearch-creating-library-linked-data-with-wikibase-project-passage.pdf.

Hegde, Medha. 2012. « Ecotones : the transitional zones ». *Biotech Articles*, n 12. http://www.biotecharticles.com/Biology-Article/Ecotones-The-Transitional-Zones-2191.html.

Hendler, Jim, et others. 2011. « Why the Semantic Web will never work ». In *7th Extended Semantic Web Conference (ESWC 2011), Crete, Greece*. http://videolectures.net/eswc2011_hendler_work/.

Internet Philosophy Ontology (InPhO) Project. s. d. « The InPhO Project ». Consulté le 19 juin 2020. https://www.inphoproject.org/.

Jaeger, Paul T, Jimmy Lin, Justin M Grimes, et Shannon N Simmons. 2009. « Where is the cloud ? Geography, economics, environment, and jurisdiction in cloud computing ». *First Monday* 14 (5).

Klein, Max. 2012. « VIAFbot Debriefing ». *OCLC Research*. https://hangingtogether.org/?p=2306.

Krafft, Dean, et Tom Cramer. 2014. « Video : Linked Data For Libraries (LD4L) Project Update ». *Coalition for Networked Information*. https://www.cni.org/news/video-linked-data-for-libraries-ld4l-project-update.

Lam, Dominic. 2014. « Big Data Challenges in Social Sciences & Humanities Research ». *Datanami*. https://www.datanami.com/2014/09/08/big-data-challenges-social-sciences-humanities-research/.

« Learn More About WorldCat Works ». 2015. *OCLC Developer Network*. https://www.oclc.org/developer/news/2014/learn-more-about-worldcat-works.en.html.

« Linked Data for Libraries (LD4L) ». 2014. https://wiki.lyrasis.org/pages/viewpage.action?pageId=41354028.

LODE : Linked Open Data Enhancer. s. d. « Gihub Linkedhumanities/lode ». Consulté le 19 juin 2020. https://github.com/linkedhumanities/lode.

McCarty, William. 2005. *Humanities Computing*. Palgrave Macmillan UK.

Nardi, Bonnie, et O'DayVicki. 1999. « Information Ecologies : Using Technology with Heart-Chapter Four ». *First Monday* 4 (5). http://firstmonday.org/ojs/index.php/fm/article/view/672/582.

OCLC Research. 2014. « Scholars' Contributions to VIAF ». https://www.oclc.org/research/areas/data-science/viaf-scholars.html.

« Open Annotation Data Model ». 2013. http://www.openannotation.org/spec/core/.

Pan-Canadian Documentary Heritage Network. s. d. « Linked Open Data (LOD) Visualization "Proof-of-Concept." ». *Canadiana*. Consulté le 13 septembre 2015. http://www.canadiana.ca/sites/pub.canadiana.ca/files/PCDHN/%20Proof-of-concept/_Final-Report-ENG/_0.pdf.

Price, Gary. 2012. « Video : "Out of the Trenches : A Linked Open Data Project" From the Pan-Canadian Documentary Heritage Network ». *LJ infoDOCKET*. https://www.infodocket.com/2012/10/25/video-out-of-the-trenches-a-linked-open-data-project-from-pan-canadian-documentary-heritage-network/.

Risser, Paul G. 1990. « The ecological importance of land-water ecotones ». In *The ecology and management of aquatic-terrestrial ecotones*, édité par H Décamps et Naiman R J, 7-21. Paris : UNESCO.

« Schema.org ». 2015. https://schema.org/.

Searle, John R. 1995. *The construction of social reality*. New York : Simon ; Schuster.

Simpson, John Edward, Susan Brown, et Lisa Goddard. 2013. « A Humanist Perspective on Building Ontologies in Theory and Practice. » In *Digital Humanities Conference Abstracts 2013*, édité par University of Nebraska, 403-5. Lincoln. http://dh2013.unl.edu/abstracts/ab-413.html.

Smith-Yoshimura, Karen, David Michelson, et Beth Mardutho. 2013. « Irreconcilable differences ? Name authority control & humanities scholarship ». *OCLC Research*. http://hangingtogether.org/?p=2621.

« The Muninn Project ». 2015. http://blog.muninn-project.org/.

The Stanford Natural Language Processing Group. s. d. « Software > Stanford Named Entity Recognizer (NER) ». Consulté le 19 juin 2020. https://nlp.stanford.edu/software/CRF-NER.html.

Uddin, Mueen, et Azizah Abdul Rahman. 2011. « Techniques to implement in green data centres to achieve energy efficiency and reduce global warming effects ». *International Journal of Global Warming* 3 (4) : 372-89.

« VIAF ». 2015. https://viaf.org/.

« VIVO Open Research Networking Community Group ». 2015. https://www.w3.org/community/vivo/.

Warren, Robert. 2012. « Creating specialized ontologies using Wikipedia : The Muninn experience ». *Proceedings of Wikipedia Academy : Research and Free Knowledge (WPAC2012)*. https://wikipedia-academy.wikimedia.de/w/images.wikipedia-academy-2012/0/0f/21_Paper_Robert_Warren.pdf.

Widmer, Rolf, Heidi Oswald-Krapf, Deepali Sinha-Khetriwal, Max Schnellmann, et Heinz Böni. 2005. « Global perspectives on e-waste ». *Environmental impact assessment review* 25 (5) : 436-58.

Wuppleman, William. 2012. « Out of the trenches : A linked open data project ». *Canadiana*. https://www.canadiana.ca.