Le web sémantique

*Analyse I/IST n°33 - Novembre 2020*

Ce devait être le web 2.0, puis le web 3.0, voire 4.0… Le web sémantique ou web des données n’en finit pas d’être la prochaine révolution sans cesse repoussée.

Et pourtant cet idéal répond en recherche à des besoins bien concrets : les bases de données scientifiques et les infrastructures de recherche souffrent d’un manque de coordination. Les données sont trop souvent recréées de zéro alors qu’il pourrait être envisageable de mutualiser des travaux antérieurs. Le développement récent de grandes bases contributives comme Wikidata, pourrait bien contribuer à diffuser à grande échelle les techniques et les normes du web sémantique au risque de changer en profondeur sa structure décentralisée.

Cette note de synthèse revient sur une révolution scientifique de longue durée qui est encore loin de s’achever.

La révolution inachevée du web sémantique

En 2001, Tim Berners-Lee annonce qu’une « nouvelle révolution va profondément affecter la diffusion de l’information scientifique ». Cette révolution s’appelle le « web sémantique » ou « web des données liées » (*linked data*) : de nouvelles infrastructures et de nouveaux langages informatiques permettent d’identifier les données contenues dans les textes.

Le web classique est en effet un web de documents plutôt qu’un web d’informations. Les métadonnées contenues dans les fichiers HTML ne décrivent que la structure éditoriale : les titres, les paragraphes, les légendes ou les signatures. Le contenu effectif du texte tels que les personnes, les lieux ou les objets mentionnés ne sont pas décrits explicitement. Si nous avions voulu sémantiser le début de ce texte nous aurions pu écrire : « En <date>2001</date>, <person page\_wikipedia=”Tim Berners-Lee” >Tim Berners-Lee</person> ».

Berners-Lee prédit qu’une transformation radicale, comparable à l’avènement du libre accès, est à l’ordre du jour et va entraîner une redéfinition complète de la notion d’article scientifique :

Quel sera l'impact du web sémantique sur l'édition scientifique ? Au cours des prochaines années, nous pouvons nous attendre à ce que les outils d'édition vont automatiquement aider les utilisateurs à inclure des marqueurs sémantiques (…) Les articles de ce type seront trouvés par de nouveaux et meilleurs moteurs de recherche, de telle sorte que les utilisateurs pourront soumettre des requêtes beaucoup plus précises. De plus, les résultats expérimentaux pourront eux-mêmes être publiés sur le web, en dehors du contexte d'un article de recherche (…) De la même manière que les preprints ont

remis en question les publications des revues scientifiques, ces nouveaux « articles in progress » vont représenter un défi significatif pour les éditeurs scientifiques[[1]](#footnote-1).

La proposition de Berners-Lee va au-delà d’un simple étiquetage des données. L’enjeu est aussi de décrire les relations qu’elles entretiennent entre elles à partir de « triplés » reliant deux entités par l’intermédiaire d’une propriété. Par exemple pour indiquer que « Tim Berners-Lee a écrit *Publishing on the semantic web* », le web des données utiliserait la syntaxe suivante : identifiant « Tim Berneers-Lee » > propriété « est auteur de » > identifiant de l’article *Publishing on the semantic web*. Cette construction s’inspire des énoncés élémentaires de la rhétorique classique du style « Socrate est un homme ». Le vocabulaire spécialisé du web va d’ailleurs très largement s’inspirer de la terminologie de la philosophie logique. Les descriptions des entités sont ainsi qualifiées, un peu pompeusement, d’*ontologies*, ou de science des êtres.

La description des relations n’est pas limitée à ce cadre élémentaire. Il est possible d’y joindre des attributs, des qualifications ou attacher une source ou une valeur temporelle à une affirmation. La force première du web sémantique vient de cette capacité à documenter non seulement des ressources mais à établir entre elles des réseaux de significations élaborés, capables de préserver, en théorie, la complexité de l’information scientifique. Les bases de données traditionnelles souffrent fréquemment d’un déficit de nuance. Typiquement dans les projets historiques portant sur des périodes un peu anciennes, les dates de naissance ou de mort sont rarement précises. Il peut aussi exister des versions discordantes du récit d’une biographie et de la datation ou de la nature des événements représentés. Les structures tabulaires trop rigides ne permettent pas de donner à lire les incertitudes et les controverses qui sont pourtant partie intégrante de la recherche scientifique.

Le web sémantique ambitionne également de surmonter l’une des principales limites de la numérisation de l’information scientifique : son isolation. La proposition de Berners-Lee repose ainsi sur un écosystème décentralisé de références croisées, qui permet à chaque acteur de se spécialiser là où il est le plus compétent et de déléguer la documentation d’entités annexes à d’autres acteurs. Chaque entité documentée est aussi une page web, baptisée URI pour *Uniform Resource Identifer* ou Identifiant de Ressource uniforme, par analogie avec l’URL.

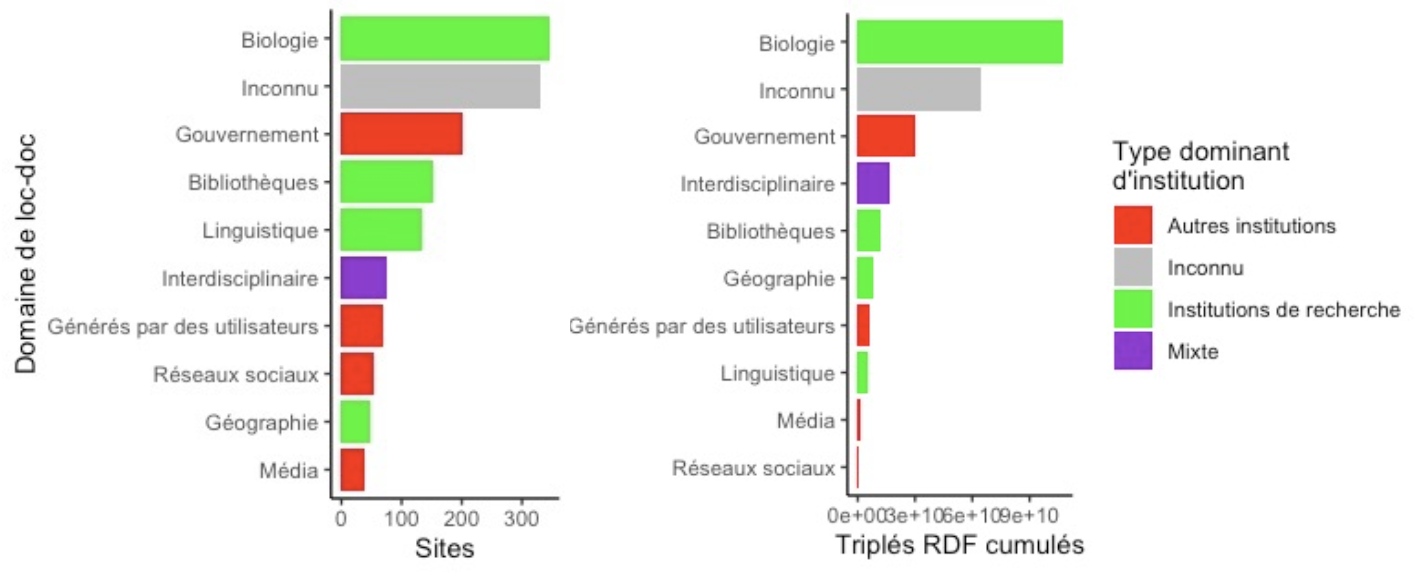
Quinze ans plus tard, la révolution du « web sémantique » reste encore à accomplir. Dans leur très grande majorité les textes scientifiques restent publiés sur un format très proche de l’imprimé (le PDF) ou dans une architecture HTML élémentaire (qui, comme pour la plupart des pages web, ne distingue que les structures éditoriales fondamentales : paragraphe, tableau, titre, etc.). Berners-Lee semble avoir sous-estimé l’attachement « culturel » des communautés scientifiques à la forme « article », que le développement des *preprints* ne

remettait pas en cause (d’autant qu’il ne faisait que généraliser des pratiques de dissémination alternative déjà très répandues).

Plusieurs utilisateurs historiques portent aujourd’hui un constat d’échec partiel. Gautier Poupeau a dressé en 2018 un bilan en demi-teinte d’une dizaine d’années d’essais à l’INA, la BNF ou dans l’infrastructure des humanités numériques Humanum. Et tout récemment, le journal de référence des bibliothèques américaines *Library Journal* vient de titrer sur *le mythe du linked data* qui est loin d’avoir accompli toutes ses promesses originelles.

L’écosystème du web sémantique aujourd’hui

Le projet lod-cloud catalogue depuis plusieurs années les principaux producteurs de données sémantisées en ligne. En 2020, la base recense 1447 sites. Elle ne prétend pas être représentative : l’inclusion se fait sur la base du dépôt volontaire, sous réserve de respecter quelques critères élémentaires (comme la présence de 1000 “triplets” de données originaux dans la base). La structure de la base est, en elle-même, évocative des limites du web sémantique. Chaque site peut être classé parmi une dizaine de domaines qui croisent en réalité plusieurs catégorisations : des disciplines (géographie, linguistique, biologie), des champs éditoriaux (*publications* [en réalité principalement des bibliothèques] et média), des domaines d’activités (gouvernement, réseaux sociaux) et le contexte de production de la base (généré par des utilisateurs). Malgré ses ambiguïtés, ce classement met en évidence le rôle essentiel joué par certaines disciplines et institutions scientifiques au sens large dans le développement du web sémantique. La visualisation ci-dessous met en évidence la répartition des domaines par nombre de sites et par “triplets”, soit approximativement par nombre d’acteurs représentés et par amplitude des informations versées à l’écosystème du web sémantique.



**Répartition des domaines scientifiques dans lod-cloud par sites et par triplets.**

Les domaines scientifiques les mieux représentés partagent des problématiques assez similaires en terme de gestions de données. Ils s’inscrivent dans une démarche de *middle big data*, avec des collectes d’informations importantes, sans être écrasantes, nécessitant potentiellement de décrire des interactions complexes. Les masses énormes de données manipulées en astronomie ou en physique des hautes énergies ne sont pas représentées : les infrastructures du web sémantique ne tiendraient pas la charge.

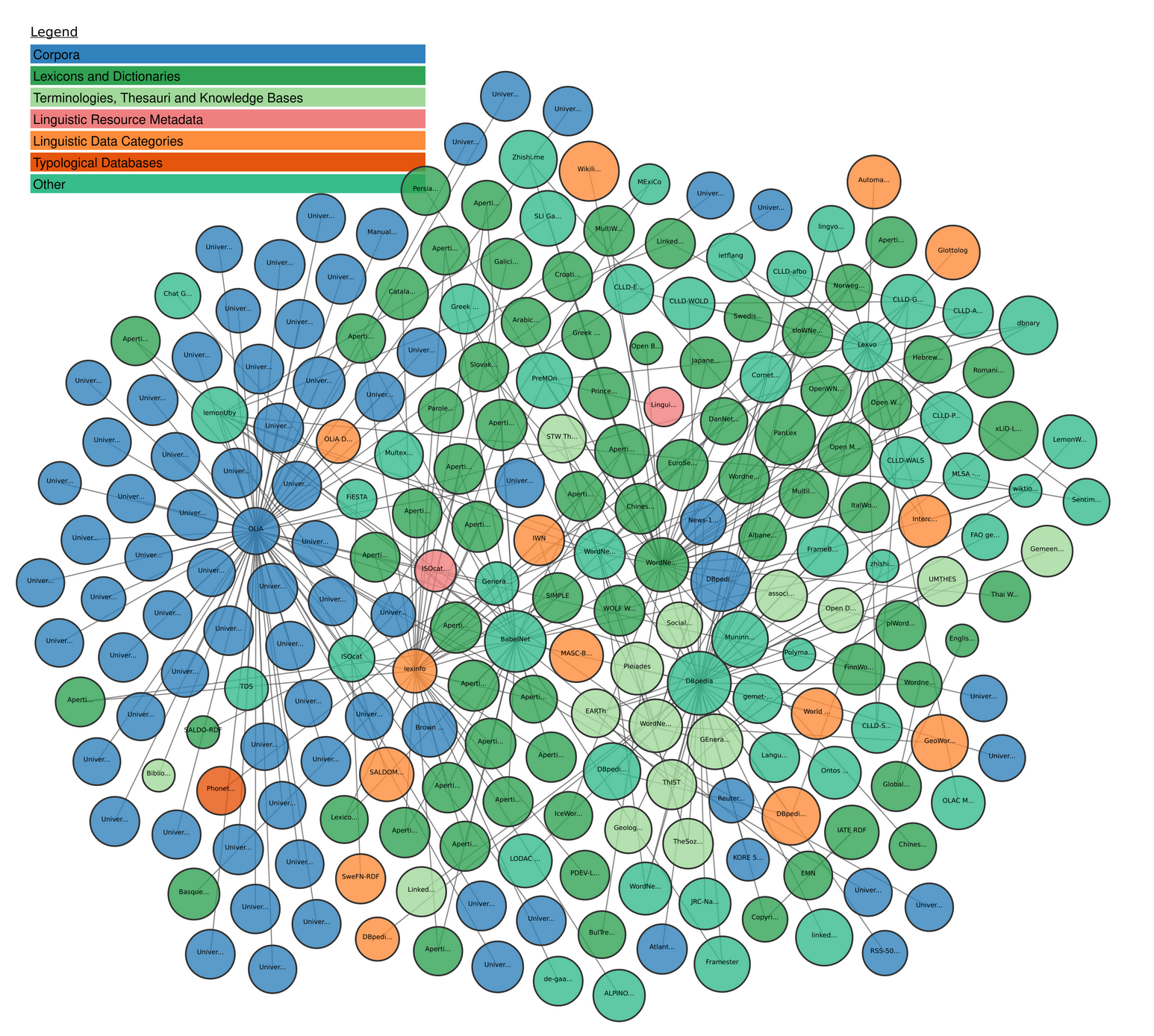
Les langages de *linked data* comme RDF ne sont pas en eux-mêmes des bases de données. Ils doivent être transformés et compilés dans de grandes bases d’énoncés ou triplets ce qui nécessite d’utiliser des outils de gestion de bases de données spécialisés comme Virtuoso. Concrètement la sophistication du web sémantique se traduit par des performances nettement plus réduites que les bases de données traditionnelles. Pour le projet Data BNF, la BNF a d’abord démultiplié les instances libres de Virtuoso pour créer un moteur de recherche suffisamment performant avant de se résoudre à acquérir une licence payante.

Pendant des années, les bibliothèques ont réduit leurs effectifs dédiés à la création de métadonnées tout en étant de plus en plus dépendant des métadonnées fournies par les éditeurs (…) Les ressources nécessaires pour maintenir les normes des données représentent une petite fraction des besoins du web sémantique[[2]](#footnote-2).

Les principaux projets en biologie se situent aujourd’hui dans la limite haute de ce qu’il est possible de faire avec ces technologies. La base universelle des protéines (*Universal Protein Resource*) est le plus gros projet indexé dans lod-cloudavec 55 milliards de triplés RDF, soit cinq fois la taille de Wikidata. L’intégration et la gestion d’une base aussi considérable ont nécessité d’importantes optimisations techniques (Bolleman et al. 2019).

En France, la pénétration du web sémantique est surtout perceptible dans un autre domaine d’activité : la documentation du patrimoine culturel et de la production scientifique. La moitié des projets recensés et, surtout, tous les acteurs importants s’inscrivent dans ce cadre : la BNF (data.bnf.fr), l’ABES (theses.fr, calames), Persée, les Archives Nationales, HAL (curieusement classé en “gouvernement”) et Isidore.

Les projets de web sémantique ne sont pas isolés : ils échangent des données les uns avec les autres dans des proportions variables. Le projet lod-cloud recense actuellement 17 000 réutilisations significatives. Sans surprise les interactions les plus importantes ont lieu entre projets biologiques (413 millions de données importées de *Gene Ontology* dans *Universal Protein Resource*) ou entre des portails bibliographiques (par exemple 137 millions de données importées de DOI dans Medline). Cette interconnexion n’est pas limitée aux projets massifs. Comme le montre le réseau suivant, les domaines de taille plus restreinte s’intègrent également dans des réseaux d’échange, ce qui matérialise la réalisation de l’une des ambitions premières du web sémantique : mettre un terme à l’isolation des bases de données scientifiques (et les duplications de travail que cela entraîne).

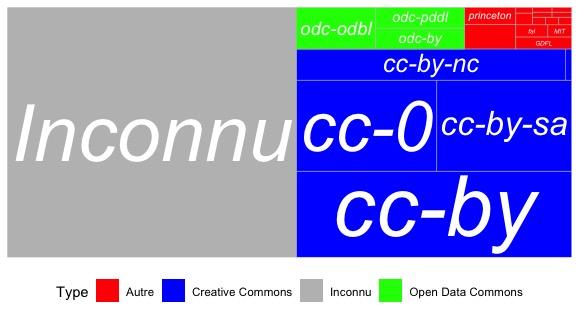


**Un domaine subsidiaire et pourtant bien connecté dans lod-cloud : la linguistique.**

L’émergence de communautés d’échange de données a été rendu possible par l’élaboration de standards communément partagés au sein d’une ou de plusieurs disciplines. Avant même le développement du web sémantique, les métiers des bibliothèques avaient commencé à créer un schéma international des données bibliographiques, Dublin Core. Ce travail de codification est également très avancé dans les sciences biologiques avec plusieurs standards couvrant les besoins précis de certains champs de recherche comme Agroportal Bioportal ou la *Wheat Data interoperability[[3]](#footnote-3)*.

Les principaux projets de web sémantique sont aujourd’hui reliés en écriture mais pas forcément en lecture. La récupération croisée des données de plusieurs bases (ou « requêtes fédérées ») reste encore problématique. Les problèmes de performance sont démultipliés par l’appel de plusieurs services n’ayant pas toujours la même capacité ou n’étant pas disponibles au même moment (Saleem et al. 2016). Même en biologie, le champ disciplinaire ayant le mieux intégré les techniques du web sémantique, les requêtes fédérées restent rares : « nous n’avons trouvé aucune application majeure qui relie différentes sources de données connectées pour produire de nouvelles connaissances (Polleres et al. 2020) ».

Au-delà des difficultés techniques, le cadre légal du *linked open data* reste très flou. La moitié des projets ne mentionnent aucune licence de réutilisation (736 sur 1447). À défaut les lois sur le droit d’auteur et le droit des bases de données s’appliquent : aucune réutilisation significative n’est légale sans l’accord explicite du producteur de la base.

**Visualisation des licences indiquées dans les projets de lod-cloud.**

Les licences les plus utilisées, les Creative Commons, ont d’abord été conçues pour des productions textuelles et visuelles, puis élargies depuis quelques années aux bases de données. Elles correspondent à un dégradé de licences, plus ou moins compatibles avec l’idéal d’un écosystème de « données liées ». La licence CC-By reste la plus employée : elle impose une contrainte d’attribution qui est généralement préférable dans un contexte scientifique où la traçabilité des données est essentielle. Plus récente, la licence CC-0 suspend simplement l’ensemble des droits de propriété intellectuelle : dans le contexte du droit des bases de données cela inclut également l’attribution, dans la mesure où le droit moral n’est pas appliqué. Enfin, les licences CC-By-SA et CC-By-NC imposent des restrictions supplémentaires: dans le premier cas le projet doit maintenir la licence des données à l’identique et dans le second cas il doit en faire un usage non-commercial. La seconde famille de licence Open Data Commons a été introduite plus tardivement pour répondre aux défaillances des licences Creative Commons en matière de bases de données. Enfin, un peu moins de 10% des licences utilisées sont rares voire spécifiques au projet.

Ces incertitudes légales compliquent la réalisation du projet initial du web sémantique. Les requêtes fédérées peuvent ainsi déboucher potentiellement sur un mille-feuille de droits de réutilisation possible allant de la propriété intellectuelle classique au domaine public en passant par différentes situations intermédiaires.

Le niveau d’intégration des 1447 bases indexées par lod-cloud reste ainsi imparfait. Et il s’agit pourtant là du cœur du web sémantique. La popularisation des standards s’effectue par bribes. Certaines normes associées comme le référentiel bibliographique Dublin Core sont intégrées d’office dans des outils majeurs de publications scientifiques, comme Omeka. En l’absence d’un important travail de fond réalisé uniquement dans certains domaines

scientifiques et dans certains milieux professionnels associés comme les bibliothèques, la production de données liées reste une vague promesse théorique. À moins que les contributions ne finissent pas converger sur un unique projet généraliste centralisé…

L’effet Wikidata : le futur du web sémantique est-il centralisé ?

Wikidata a longtemps été un serpent de mer. Le projet est officiellement lancé en 2013 mais il apparaît dix ans plus tôt dans les discussions internes des communautés de Wikipédia. Dès 2004, il apparaît nécessaire de mutualiser les efforts de documentation entrepris par les différentes versions linguistiques de Wikipédia : documenter « tous les médias (photos, sons, vidéos, etc.) sur un seul site est une nécessité (plusieurs fois évoqué sur les listes de discussions). Mais je crains que ça ne se fasse pas à court terme. Ceci dit, ce site (Wikidata?) ne serait que pour stocker des données[[4]](#footnote-4). »

Quinze ans plus tard le bilan de Wikidata est surprenant.

Le site n’est pas parvenu à accomplir son ambition première : mutualiser les données de Wikipédia et d’autres projets-frères. Dans Wikipédia la réutilisation des informations est généralement limitées aux « infobox », les petites fiches sur le côté résumant les principales données factuelles du sujet d’un article, et, même là, elle ne fait pas l’unanimité.

**Exemple d’infobox : la page Wikipédia d’Open Edition**

Pour les contributeurs réguliers de Wikipédia, l’importation automatique des données depuis Wikidata ajoute une couche de complexité : elle nécessite de s’initier à un projet très différent et, à ce jour, il n’est pas possible de suivre à distance les modifications effectuées sur Wikidata (comme on le fait communément pour les pages de Wikipédia avec la « liste de suivi »). Des versions linguistiques de petites tailles de Wikipedia néanmoins ont commencé à utiliser Wikidata pour générer automatiquement des articles avec des résultats contrastés : la communauté n’est pas toujours en mesure de gérer un stock d’articles trop important (notamment pour éviter le vandalisme)[[5]](#footnote-5). Un nouveau projet, *Abstract Wikipedia* a été récemment accepté pour améliorer le rendu des articles générés à partir d’une grammaire

« abstraite » qui pourrait être restituée dans n’importe quelle langue[[6]](#footnote-6).

Et pourtant, Wikidata a eu un succès imprévu auprès des institutions scientifiques et des bibliothèques. Le site commence à s’imposer comme une *méta-référence* par défaut, fédérant les principaux identifiants et permettant de naviguer d’une base à l’autre. Son impact est particulièrement important dans des disciplines jusqu’ici peu impliquées dans le développement du web sémantique comme l’histoire ou les humanités. Des projets de prosopographies sont ainsi menés sur des sujets aussi divers que la réforme protestante (“Mapping the Scottish Reformation:” n.d.), la vie de Dante Alighieri (“Populating Narratives Using Wikidata Events: An Initial Experiment” n.d.), la peinture hollandaise, les procès en sorcellerie[[7]](#footnote-7) ou les structures de l’ESR en France[[8]](#footnote-8). Wikidata a commencé à agréger massivement certaines bases préexistantes comme les références bibliographiques, sous l’impulsion des projets Scholia et Wikicite : en février 2020, près d’un tiers des items de la base sont des articles scientifiques[[9]](#footnote-9).

Wikidata a le mérite de régler par anticipation de nombreuses questions techniques liées à la gestion et la récupération des données, ce qui permet à des projets de petite taille de se focaliser sur la contribution. La structure de la base se prête bien à des données parfois complexes ou lacunaires. Il peut exister plusieurs versions concurrentes d’une même déclaration comme, par exemple, plusieurs dates de naissance s’il n’existe pas de consensus pour une date précise (Voß 2016). Les déclarations peuvent être précisées par des “qualificateurs” qui indiquent, par exemple, qu’une information n’est valide que pour une certaine période. Enfin, chaque déclaration doit, idéalement, être référencée. Cet impératif de citation des sources, originellement importé de Wikipédia, contribue à changer profondément les pratiques épistémologiques du web sémantique et, par là, les structures

éditoriales : la donnée, comme le texte, devient un objet référençable et traçable. Avec l’inclusion des sources et des qualificateurs, Wikidata n’est plus réellement conforme au standard RDF : le site a graduellement créé une nouvelle norme de représentation de données.

À la différence de ses prédécesseurs, Wikidata dispose non seulement d’une infrastructure technique permettant de formuler des requêtes complexes sur un vaste corpus mais aussi d’une communauté très active capable d’améliorer constamment l’accessibilité et les fonctionnalités des outils de récupération des données. Le point d’accès SPARQL est ainsi associé à tout un répertoire de datavisualisations permettant de projeter les données en temps réel sous forme de cartes ou de graphes interactifs.

Le développement de Wikidata a permis de surmonter plusieurs limites majeures du web sémantique au risque de le changer en profondeur. Devenu une référence par défaut, le site tend à centraliser l’échange des données. Pour l’instant ce basculement n’est pas encore perceptible dans la cartographie de lod-cloud, en partie à cause du poids encore considérable des bases en biologie. Néanmoins il est nettement ressenti par plusieurs observateurs externes :

D’un hub d’identifiants/un référentiel, Wikidata tend peu à peu par simplicité technique là aussi à devenir la base où sont stockées toutes les données. Pourquoi s’embêter à maintenir une exposition des données en Linked Data sur sa propre infrastructure si toutes les données sont par ailleurs disponibles dans Wikidata ? Nous avons là encore un bel exemple d’une dérive centralisatrice à partir d’une idée initiale décentralisatrice[[10]](#footnote-10).

La centralisation opérée par Wikidata n’est cependant pas du même ressort que celle mise en œuvre par les grands acteurs commerciaux du web. Le site reste participatif et dans sa mise en œuvre, il s’apparente plutôt à une constellation de projets thématiques, les *Wikiprojects*, qui jouent un rôle déterminant dans la création, la définition et l’adoption des propriétés. La communauté de Wikidata n’est d’ailleurs elle-même pas favorable à une centralisation trop poussée qui aurait pour effet de noyer le projet et sa cohérence éditoriale sous une masse énorme de données trop spécialisées. Les projets d’intégration en masse comme *Wikicite*, suscitent beaucoup de débats au sein de la communauté. Plusieurs scénarios sont actuellement envisagés, de l’intégration complète des bases spécialisées dans un Wikidata 10 à 20 fois plus gros à la fédération de bases décentralisées, en passant par la constitution de projets-frères[[11]](#footnote-11). Les enjeux techniques, comme la faisabilité des requêtes fédérées sur SPARQL, sont ici aussi déterminants que les questions de gouvernance et d’identité des projets.

En 2020, le web sémantique reste avant tout un « projet en construction » (Gandon 2018) qui s’est déjà solidement implanté dans certaines disciplines et commence à impliquer de nouvelles communautés de recherche. L’émergence de portails généralistes comme Wikidata contribue à résoudre par anticipation de nombreux défis techniques tout en reposant frontalement la question du devenir de l’écosystème des données liées.

Références

Bolleman, Jerven, Eduoard de Castro, Delphine Baratin, Sebastien Gehant, Beatrice A. Cuche, Andrea H. Auchincloss, Elisabeth Coudert, et al. 2019. “HAMAP Rules as SPARQL A Portable Annotation Pipeline for Genomes and Proteomes.” Preprint. Bioinformatics. <https://doi.org/10.1101/615294>.

Gandon, Fabien. 2018. “A Survey of the First 20 Years of Research on Semantic Web and Linked Data.” Ingénierie Des Systèmes d’information 23 (3–4): 11–38. <https://doi.org/10.3166/isi.23.3-4.11-38>.

“Mapping the Scottish Reformation:” n.d. Accessed September 17, 2020. <https://doi.org/10.21083/irss.v44i0.5834>.

Polleres, Axel, Maulik Rajendra Kamdar, Javier David Fernández, Tania Tudorache, and Mark Alan Musen. 2020. “A More Decentralized Vision for Linked Data.” Semantic Web 11 (1): 101–13. <https://doi.org/10.3233/SW-190380>.

“Populating Narratives Using Wikidata Events: An Initial Experiment.” n.d. Accessed September 17, 2020. <https://doi.org/10.1007/978-3-030-11226-4_13>.

Saleem, Muhammad, Yasar Khan, Ali Hasnain, Ivan Ermilov, and Axel-Cyrille Ngonga Ngomo. 2016. “A Fine-Grained Evaluation of SPARQL Endpoint Federation Systems.” Semantic Web 7 (5): 493–518. <https://doi.org/10.3233/SW-150186>.

Voß, J. 2016. “Classification of Knowledge Organization Systems with Wikidata.” In NKOS@TPDL. <https://doi.org/10.5281/ZENODO.61767>.

1. Nous traduisons : Tim Berners-Lee et James Hendler, « Publishing on the semantic web », *Nature*, 410(6832), 2001, p. 1023‑1024. [↑](#footnote-ref-1)
2. Nous traduisons : Tim Berners-Lee et James Hendler, « Publishing on the semantic web », *Nature*, 410(6832), 2001, p. 1023‑1024. [↑](#footnote-ref-2)
3. <https://www.rd-alliance.org/groups/agriculture-data-interest-group-igad/wiki/working-group-%E2%80%9Cwheat-data-interoperability%E2%80%9D.html> [↑](#footnote-ref-3)
4. Contribution d’Aoineko du 3 août 2004 : <https://fr.wikipedia.org/wiki/Wikip%C3%A9dia:La_cave_du_Bistro#Sous-section_pour_la_commodit%C3%A9_de_l'%C3%A9dition> [↑](#footnote-ref-4)
5. Le cas le plus extrême est celui de la Wikipédia en Cebuano, une langue des Philippines : il s’agit aujourd’hui de la deuxième Wikipédia par nombre d’articles (5,3 millions). L’essentiel du contenu a été produit par un robot, Lsjbot, [↑](#footnote-ref-5)
6. <https://meta.wikimedia.org/wiki/Abstract_Wikipedia> [↑](#footnote-ref-6)
7. <https://thinking.is.ed.ac.uk/wikidata-workshop/real-world-datasets-the-survey-of-scottish-witchcraft/> [↑](#footnote-ref-7)
8. <https://twitter.com/JulienGossa/status/1153752700552716288> [↑](#footnote-ref-8)
9. <https://www.wikidata.org/wiki/Wikidata:Statistics> [↑](#footnote-ref-9)
10. <http://www.lespetitescases.net/reflexions-et-questions-autour-du-web-semantique> [↑](#footnote-ref-10)
11. <https://www.wikidata.org/wiki/Wikidata:WikiCite/Roadmap#Four_scenarios_for_the_future_of_WikiCite> [↑](#footnote-ref-11)