Mémoire d' étude / Mars 2020

Diplôme de conservateur de bibliothèque

OAI-PMH à « l'heure du web sémantique » : bilans et perspectives

Vincent de Lavenne de la Montoise

Sous la direction d'Anna Svenbro
Responsable de l'informatique documentaire – BU de Limoges

Remerciements

Merci à Anna Svenbro, de m'avoir accompagné dans ce beau défi : faire un mémoire entier sur un protocole informatique dont le seul nom provoque des émotions allant de l'effroi à l'ennui le plus total. Merci pour la patience, les discussions, les relectures attentives qui m'ont ouvert des horizons inattendus.

Merci à Nola N'Diaye, qui m'a aidé à comprendre tant de choses qui se retrouveront, je l'espère, dans ce mémoire, qui a formulé, reformulé et re-re-formulé certaines choses pour que je les comprenne, et qui, en plus, a su me diriger vers les personnes qui ont nourri ma réflexion. Merci aussi à Isabelle Bizos, qui m'a expliqué un autre contexte, et les défis qui se posaient.

Merci à toute la constellation des DCB 28 – promotion Louise Michel, qui ont écouté avec patience et attention mes élucubrations sur les données, supporté mes mauvais jeux de mots sur le sujet et bien plus encore.

Merci à tous les collègues de McGill qui ont supporté mes questionnements incessants, et surtout Megan Chellew, Robin Desmeules et Anna Dysert.

Merci aussi à Mélanie, Stéphanie, Safia, Philémon, Richard et Marie-Claire, Serge, Jim et Bertrand, Gerry, Michel et Marc, Gerry, et aussi Lucy, Molly, Matthew, George, Laura, Jason, Jerry de m'avoir accompagné pendant un moment ou un autre de ce travail

Merci enfin à Sophie, pour tout.

Résumé : À l'approche du vingtième anniversaire du protocole OAI-PMH, et dans un environnement web qui a subi de profondes évolutions (technologiques et d'usages), quelle est l'actualité de l'échange de données ? Comment se sont construits les usages des professionnel-le-s en la matière ? Sont-ils adaptés aux défis actuels ? Ce travail se propose d'analyser l'exposition et l'échange de données sous un angle historique, avant d'essayer de comprendre les enjeux actuels qui détermineront quelle(s) solution(s) techniques choisir.

Descripteurs : Web sémantique, Web 2.0, Édition en libre accès, Archives ouvertes de publications scientifiques, Métadonnées, Bibliothèques numériques

Abstract : While we are nearing the twentieth birthday of the OAI-PMH protocol, and in a web environment which went through massive changes, both in technology and in usage, what is the actuality of data exchange ? How did the community build its practices? Are they still relevant in the light of today's challenges? This work analyses data exposition and exchange from a historical standpoint, before trying to understand the current challenges which will determine what technical solution(s) should be used.

Keywords : Semantic web, Web 2.0, OAI-PMH, OAI-ORE, Open access publishing, Archives, Metadata, Metadata harvesting, Digital libraries

Droits d'auteurs

-

SOMMAIRE

SIGLES ET ABREVIATIONS

ABES : Agence Bibliographique de l'Enseignement Supérieur

API : Application Programming Interface

BnF : Bibliothèque nationale de France

CCSD : Centre pour la Communication Scientifique Directe

CINES : Centre Informatique National de l'Enseignement Supérieur

COAR : Confederation of Open Access Repositories

DOAJ : Directory of Open Access Journals

EDM : Europeana Data Model

ENSSIB : École Nationale Supérieure des Sciences de l'Information et des Bibliothèques

ESE : Europeana Semantic Elements

ETD : Electronic Theses and Dissertations

FOAF : Friend Of A Friend

FRBR : Functional Requirements for Bibliographic Records

FTP : File Transfer Protocol

GLAM : Galleries, Libraries, Archives and Museums

HAL : Hyper Articles en Ligne

HTML : Hypertext Markup Language

HTTP : Hypertext Transfer Protocol

IdHAL : Identifiants Hyper Articles en Ligne

IdRef : Identifiants et Référentiels

IFLA : International Federation of Library Associations

IFLA-LRM : IFLA Library Reference Model

IIIF : International Image Interoperability Framework

LANL : Los Angeles National Laboratory

JSON-LD : JavaScript Object Notation for Linked Data

LOM : Learning Object Metadata

MARC : Machine-Readable Cataloging

METS : Metadata Encoding and Transmission Standard

OAI : Open Archives Initiative

OAI-PMH : Open Archives Initiative-Protocol for Metadata Harvesting

OAI-ORE : Open Archives Initiative-Object Reuse and Exchange

ORCID : Open Researcher and Contributor Identifier

ORI-OAI : Outil de Référencement et d'Indexation-Open Archives Initiative

RDA : Resource Description and Access

RDF : Resource Description Framework

RePEc : Research Papers in Economics

RSS : Really Simple Syndication

SciELO : Scientific Electronic Library Online

SIGB : Système Intégré de Gestion de Bibliothèque

SKOS : Simple Knowledge Organization System

SPAR : Système de Préservation et d'Archivage Réparti

SPARC : Scholarly Publishing and Academic Resources Coalition

SPARQL : SPARQL Protocol And RDF Query Language

SRU : Search/Retrieve via URL

SRW : Search/Retrieve Web Service

TEF : Thèses Électroniques Françaises

URI : Uniform Resource Identifier

URL : Uniform Resource Locator

UNT : Université Numérique Thématique

W3C : World Wide Web Consortium

XML : Extensible Markup Language

INTRODUCTION

"Bad Libraries build collections. Good libraries build services. Great libraries build Communities"

R. David Lankes

De la description à la diffusion

À quoi sert une bibliothèque ? Si l'on reprend la phrase fameuse de David Lankes, une bonne bibliothèque ne construit pas de collections ni de services, mais une communauté ; ou plutôt, une excellente bibliothèque construit des collections, qui lui permettent d'offrir des services qui aident à construire une communauté. La construction des collections est donc un élément primordial, le premier bloc de la construction « identitaire » d'une bibliothèque, même si ce bloc ne peut être le seul. Qu'elles soient physiques ou numériques, universelles ou partielles, incontournables ou invisibles, elles sont au cœur de l'action des bibliothécaires.

Une partie importante de la construction du travail sur une collection réside dans sa description. Pour fournir un document, il convient de savoir le décrire et d'en noter la possession. Avec des tablettes, des rouleaux, des ouvrages, des fiches papiers manuscrites ou tapées, avec un certain degré d'automatisation, puis des ordinateurs et des systèmes connectés[1]...la description des documents, le catalogage, sont devenus en partie un processus coopératif à l'échelle nationale ou internationale. En outre, l'échange des données descriptives constitue aujourd'hui un enjeu crucial. Quand nous parlons de données descriptives, nous utilisons le terme de métadonnées, soit, selon la norme ISO 19115[2], des « données qui définissent et décrivent d'autres données », or elles peuvent être descriptives, mais aussi administratives ou structurelles, et dans tous les cas, permettent de signaler et de

[1]À ce sujet, consulter ROCHE, Mélanie. *En attendant « le jour […] où il n'y aura plus de catalogue à faire » : une histoire matérielle des catalogues de bibliothèque (1789 – 1993)* [en ligne]. Mémoire d'étude. [S. l.] : ENSSIB, janvier 2014. [Consulté le 24 janvier 2020]. Disponible à l'adresse : https://www.enssib.fr/bibliotheque-numerique/documents/64118-en-attendant-le-jour-ou-il-n-y-aura-plus-de-catalogue-a-faire-une-histoire-materielle-des-catalogues-de-bibliotheque-1789-1993.pdf.

[2]INTERNATIONAL ORGANIZATION FOR STANDARDIZATION. ISO 19115-1:2014. Dans : *ISO* [en ligne]. [s. d.]. [Consulté le 24 janvier 2020]. Disponible à l'adresse : http://www.iso.org/cms/render/live/en/sites/isoorg/contents/data/standard/05/37/53798.html.

gérer l'information numérique, en documentant sa création, sa structuration, sa diffusion et ses conditions d'utilisation. Produire des métadonnées permet de rendre des objets numériques (qu'ils le soient nativement ou pas) accessibles par des humains, et lisibles par des machines. Elles peuvent apparaître de nombreuses manières : externes au document, par exemple dans un fichier l'accompagnant, encapsulées dans le document ou encore englobantes, c'est-à-dire dans une déclaration débutant le document.

Échanger ces métadonnées, c'est éviter de multiplier le travail de description par le nombre d'établissements. C'est aussi contribuer à créer une information cohérente, à permettre à chacun-e d'accéder à l'information par les mêmes biais. C'est enfin, avec le numérique, permettre de signaler des documents distants. L'échange et le partage de métadonnées, qui seul donne un sens à ce côté coopératif, sont aujourd'hui primordiaux, particulièrement dans le domaine des sciences, qui se nourrit d'articles, de livres et autres publications en permanence, avec parfois un impératif de temps (les contenus devenant périmés ou obsolètes dans un cycle assez court dans certaines disciplines), ou un obstacle de coût, lié à l'augmentation des coûts de la documentation scientifique électronique. Ainsi, l'échange des métadonnées est la clé de voûte du développement des archives ouvertes.

Faire parler les archives pour améliorer la communication scientifique

Ces dernières se mettent en place dès le début des années 90, généralement concernées seulement par une discipline, comme ArXiv pour la physique ou RePEc pour l'économie, ou propres à une langue et une zone d'influence, comme SciELO en Amérique du Sud. Indépendantes les unes des autres, et n'utilisant pas - au départ - de manière cohérente les mêmes formats et normes, ce qui crée un nouvel obstacle à la communication scientifique, celui de l'interopérabilité : comment faire communiquer des structures entre elles, indiquer des ressources dans des formats différents, lister un ensemble de documents qui sont disponibles librement dans ces archives ouvertes nouvellement créées ? Comment faire en sorte que ces nouveaux lieux de savoir « discutent » entre eux, et les lier avec d'autres institutions dépositaires de ce savoir, au premier rang desquelles les bibliothèques (mais aussi les archives et musées), dont bien des usager-e-s peuvent être intéressés par leurs contenus ? Ces questions font surgir un nouveau besoin : l'interopérabilité, c'est à dire « la capacité que possède un produit ou un système, dont les interfaces sont intégralement connues, à fonctionner avec d'autres produits ou systèmes existants ou futurs et ce sans

restriction d'accès ou de mise en œuvre »[3]. L'objectif visé n'est donc pas celui de la compatibilité, ni de la portabilité, mais bien la capacité de fonctionner, actuellement ou ultérieurement, avec d'autres formats, d'autres structures, c'est-à-dire l'interopérabilité.

C'est pour répondre à ce besoin que l'Open Archive Intiative naît à la fin des années 1990. La conférence de Santa Fe[4] réunit la communauté de professionnels de l'information, engagés dans le domaine de l'accessibilité de l'information scientifique, pour définir une manière normalisée d'exposer et d'échanger des métadonnées. En simplifiant et en normalisant les échanges de données, une nouvelle manière de rendre visibles des travaux de recherche apparaît, ce qui conserve la pertinence du mouvement des archives ouvertes et permet leur évolution et la naissance de la « voie verte » au sein du mouvement pour l'accès ouvert. La simplicité, l'élasticité du protocole facilite ensuite son adoption au sein d'autres communautés documentaires, dont celles du patrimoine ; dans un moment de coopération internationale des institutions, le besoin d'échanger des données est fort, et le protocole permet de recentraliser une information dispersée sans faire disparaître leur lieu d'exposition originale.

Un essaimage transdisciplinaire et international

En-dehors même du monde de la recherche, le protocole essaime, puisqu'il répond à un besoin général, que partage par exemple la communauté patrimoniale : rendre visible des contenus de tous types sur le web. C'est ainsi qu'OAI-PMH devient, particulièrement dans le monde des bibliothèques, un protocole familier à tou-te-s les professionnel-le-s qui gèrent des bibliothèques numériques, au-delà même du monde scientifique ; un standard qui sort de son contexte initial de création, et permet de créer des liens entre plusieurs communautés professionnelles, comme les galeries, bibliothèques, archives et musées (que l'on regroupe sous l'acronyme GLAM). Enfin, OAI-PMH est la technologie de base de projets d'envergure, tels que la bibliothèque numérique Europeana.

[3] *Définition : Interopérabilité* [en ligne]. [s. d.]. [Consulté le 24 janvier 2020]. Disponible à l'adresse : http://definition-interoperabilite.info/.

[4] VAN DE SOMPEL, Herbert et LAGOZE, Carl. The Santa Fe Convention of the Open Archives Initiative. *D-Lib Magazine* [en ligne]. Février 2000, Vol. 6, nº 2. [Consulté le 13 janvier 2020]. DOI 10.1045/february2000-vandesompel-oai.

La révolution aura-t-elle lieu ?

Or, près de 20 ans après l'apparition de ce protocole, force est de constater que le contexte est totalement différent. Les technologies du web 2.0 ont modifié les habitudes et les attentes des publics, l'émergence du « web sémantique », longtemps annoncée, promet également de bouleverser les usages La transformation du paysage éditorial scientifique contribue également à faire penser qu'OAI-PMH devrait être promis à une diminution d'usage, une disparition progressive au profit d'autres dispositifs plus récents, peut-être mieux adaptés aux nouveaux besoins de la recherche et à l'évolution technologique. Pourtant, à l'heure actuelle, il reste le vecteur principal de la transmission de métadonnées dans les institutions culturelles, même si on peut questionner sa pertinence et sa portée, en terme d'augmentation de la visibilité, de liens avec les pratiques effectives du public, en particulier du monde de la recherche. Dans un moment où les usagers du web sont autant des contributeur-rice-s que des consommateur-rice-s, à la fois amateur-ice-s et expert-e-s[5], quelle est la pertinence d'un protocole où des professionnel-le-s contrôlent l'information et sa dissémination ? Ces interrogations nous amènent à nous intéresser sur le rôle du protocole OAI-PMH en 2020, à la fois central au paysage de la bibliothèque numérique et supposément inadapté à son temps, ou au moins à son public-cible.

Mais il est également intéressant de penser aux technologies qui devaient rendre obsolètes des dispositifs tels qu'OAI-PMH et révolutionner nos pratiques : celles du web sémantique. Celui-ci, selon le Ministère de la Culture et de la Communication, « consiste [...] à interconnecter d'immenses référentiels (terminologies, catalogues d'œuvres…) ouvrant la voie à des usages radicalement nouveaux des données numériques »[6]. La promesse d'une révolution dans la manière d'aborder l'information est donc fondamentale dans le web sémantique. La non-disparition du protocole est donc, sans doute, un signe de la non-apparition de solutions qui devaient être la prochaine étape de l'organisation et l'accès de l'information dans un environnement numérique. Alors que le cadre de ces technologies a été défini dès le début des années 2000, soit dans le même temps qu'OAI-PMH, leur application n'est, à l'heure actuelle, pas à la hauteur des ambitions affichées. L'explosion de

[5] Dans l'esprit de la logique professionnel amateur, ou pro-am. À ce sujet, voir ADENOT, Pauline. Les pro-am de la vulgarisation scientifique : de la co-construction de l'ethos de l'expert en régime numérique. *Itinéraires. Littérature, textes, cultures* [en ligne]. Juin 2016, nº 2015-3. [Consulté le 27 février 2020]. DOI 10.4000/itineraires.3013.

[6] *Web sémantique, web de données, liage de données* [en ligne]. [s. d.]. [Consulté le 24 janvier 2020]. Disponible à l'adresse : https://www.culture.gouv.fr/Sites-thematiques/Langue-francaise-et-langues-de-France/Politiques-de-la-langue/Langues-et-numerique/Web-semantique-web-de-donnees-liage-de-donnees.

l'aspect social du web, la complexité des technologies sémantiques et de leur adoption, l'apparition de l'intelligence artificielle et des promesses qu'elle porte ont grandement limité la « révolution » qui était promise. Dans ce contexte, il faut sans doute ré-évaluer des dispositifs, certes plus anciens, mais dont la profession s'est pleinement emparée... Pour toutes ces raisons, il est nécessaire de se demander quelles sont les raisons pour lesquelles le protocole OAI-PMH est encore pertinent aujourd'hui. Des promesses et objectifs qui existaient à son lancement, lesquels peuvent être considérés réalisés ? Quels changements les développements du web et du paysage documentaire induisent-ils dans nos pratiques ? Est-il possible, et souhaitable, de détrôner ce protocole pour utiliser une technologie différente ? Au moment d'un mouvement institutionnel en faveur de l'accès ouvert, quelle place et quel futur pour les archives ouvertes ?

Avant de répondre à ces questions, il convient de revenir sur les ambitions de ce travail, et avant cela, de dire ce qu'il n'est pas. Ainsi, dans ces pages, on ne trouvera pas un guide d'implémentation du protocole OAI-PMH en bibliothèque, ni une simple description du protocole : des documents[7], nombreux et précis, existent pour répondre à ces besoins. De la même manière, on ne trouvera pas ici un catalogue comparatif de solutions techniques qui pourraient remplacer ce protocole. Ce mémoire se propose de tenter comprendre les enjeux de sa création, de sa diffusion, et de tenter d'avoir une vision prospective sur les échanges de données en bibliothèque à l'heure actuelle. Une consultation de la littérature nous montre que le sujet a été abondamment abordé en France, par des structures comme l'Agence Bibliographique de l'Enseignement Supérieur[8], le Ministère de la Culture et de la Communication[9], mais aussi des professionnel-le-s s'exprimant par le biais de blogs[10]. En consultant ces ressources, mais aussi la littérature internationale, en rencontrant des personnes qui utilisent le protocole, qui étudient d'autres modèles (et notamment les

[7]Par exemple, HUNTER, Philip. *OAI and OAI-PMH for absolute beginners: a non-technical introduction [Introduction to OAI and Harvesting]* [en ligne]. 2005. [Consulté le 24 juillet 2019]. Disponible à l'adresse : http://eprints.rclis.org/7143/ ; *Le fonctionnement de base (OAI-PMH) - Wiki_AO* [en ligne]. [s. d.]. [Consulté le 31 janvier 2020]. Disponible à l'adresse : http://urfist-apps.unice.fr/wiki_AO/index.php/Le_fonctionnement_de_base_(OAI-PMH) ; LAGOZE, Carl, VAN DE SOMPEL, Herbert, NELSON, Michael, et al. *Open Archives Initiative - Protocol for Metadata Harvesting - Guidelines for Repository Implementers* [en ligne]. 19 janvier 2005. [Consulté le 31 janvier 2020]. Disponible à l'adresse : https://www.openarchives.org/OAI/2.0/guidelines-repository.htm.

[8]Calames : nouveau service OAI-PMH. Dans : *FIL'ABES* [en ligne]. 10 octobre 2017. [Consulté le 31 janvier 2020]. Disponible à l'adresse : https://fil.abes.fr/2017/10/10/calames-nouveau-service-oai-pmh/.

[9]NAWROCKI, François. *Patrimoine écrit - Le protocole OAI-PMH* [en ligne]. 2 mai 2013. [Consulté le 28 juin 2019]. Disponible à l'adresse : https://web.archive.org/web/20130502181820/http://www.patrimoineecrit.culture.gouv.fr/Num/OAI-PMH.html.

[10]Par exemple : BERMÈS, Emmanuelle. Tout sur l'OAI. Dans : *Figoblog* [en ligne]. 16 février 2005. [Consulté le 24 juillet 2019]. Disponible à l'adresse : https://figoblog.org/2005/02/16/566/ ; POUPEAU, Gautier. *Les carcans de la pensée hiérarchique et documentaire (2) | Les petites cases* [en ligne]. [s. d.]. [Consulté le 31 janvier 2020]. Disponible à l'adresse : https://www.lespetitescases.net/carcans-de-la-pensee-hierarchique-et-documentaire-2.

technologies du web sémantique), ce mémoire tentera de comprendre la manière dont le protocole est utilisé aujourd'hui et de voir les possibilités qui sont offertes aux bibliothèques.

Afin de saisir le sujet dans son ensemble, il convient de revenir, dans une première partie, sur la naissance du protocole OAI-PMH : comprendre les acteurs, ainsi que le contexte de création et de dissémination permet d'expliquer le succès aussi bien que les obstacles qui y sont liés ; il faut ensuite, en seconde partie, analyser ce qu'on appelle le web sémantique, pour en comprendre le principe, les promesses et réalisations, avant d'explorer dans une troisième partie les pistes de changements qui s'offrent aux professionnel-le-s de l'information, entre solutions technologiques nouvelles et approfondissement concerté de l'utilisation de solutions existantes.

D'UNE ARCHIVE UNIVERSELLE A UN UNIVERS D'ARCHIVES : LA NAISSANCE D'UN PROTOCOLE

La conférence de Santa Fe : besoins, solutions et compromis

En 2020, le protocole OAI-PMH est la brique standard de l'implémentation d'une bibliothèque numérique. Son installation est gérée par la plupart des logiciels de bibliothèques numériques et les créateurs de Systèmes Intégrés de Gestion de Bibliothèque, comme Omeka[11] ou Koha[12]. C'est également la technologie qui permet de faire fonctionner de grands projets numériques comme Gallica ou Europeana. C'est enfin la principale source de métadonnées pour des aggrégateurs de contenus comme Isidore, ou des archives ouvertes comme, en France, la plate-forme Hyper Articles en Ligne (HAL), ou encore ArXiv. C'est donc une technologie présente aussi bien dans le domaine scientifique, toutes disciplines confondues, que dans celui du patrimoine. Mais comment s'est développé ce protocole ?

Ouvrir les archives : aux origines de l'open access

Pour comprendre, il convient de remonter aux origines du mouvement des archives ouvertes. Vers la fin des années 1980, l'utilisation croissante de l'informatique dans les travaux de recherche permet d'accélérer le rythme de la communication scientifique, en partageant directement ses travaux avec les collègues de la discipline, via le courrier électronique qui se développe durant ces années[13]. Le besoin de communication directe permet alors de contourner les délais, parfois longs, de publication dans des revues scientifiques, avec une évaluation par les pairs. Mais le nombre de personnes à contacter en même temps, le nombre de travaux de recherche à envoyer et recevoir rendent complexes ces échanges entre chercheurs. De plus, les capacités de stockage sont alors assez limitées : être dans une telle boucle de courriels, à l'époque, c'est recevoir tous les articles qui circulaient à l'intérieur. Or, plus le mouvement s'amplifie, plus la solution des échanges directs devient problématique : c'est ce qui pousse à la création d'arXiv, à Los Alamos en 1991. D'abord

[11] *OAI-PMH Harvester* [en ligne]. [s. d.]. [Consulté le 11 janvier 2020]. Disponible à l'adresse : https://info.omeka.net/build-a-website/manage-plugins/oai-pmh-harvester/.

[12] *Services Web — Documentation Koha Manual 18.11* [en ligne]. [s. d.]. [Consulté le 31 janvier 2020]. Disponible à l'adresse : https://koha-community.org/manual/18.11/fr/html/webservices.html#oai-pmh.

[13] Notamment avec la création du Simple Mail Transfer Protocol en 1981, qui contribue à l'essor des communications informatiques, en particulier dans le monde de la recherche, puis le protocole POP en 1984, et le protocole IMAP

une simple boîte mail accessible depuis n'importe quel ordinateur, puis un entrepôt électronique accessible via le File Transfer Protocol (FTP), puis enfin relié au web en 1993, arXiv est l'archive ouverte pionnière, propre – à ses débuts – au monde de la recherche en physique, où l'importance de la communication se fait particulièrement ressentir : c'est pour cela qu'elle répertorie des articles avant leur publication dans des revues (*preprints*).

Mais arXiv n'est pas la seule archive ouverte à apparaître dans les années 1990 : dans le domaine des sciences économiques, NetEc est créé, avec plusieurs facettes, pour donner accès à des *preprints*[14], des informations sur l'économie, des répertoires de pages web d'économistes[15], des logiciels et même des blagues[16] (JokEc). Ce dernier point n'est pas si anodin, puisqu'il montre le goût des chercheurs et chercheuses pour une dimension sociale et informelle, même au sein de la communication scientifique, sur laquelle nous reviendrons ultérieurement. REsearch Papers in Economy (RePEc) remplacera ensuite NetEc, qui devient rapidement une archive centrale à la communauté économique.

Le mouvement des archives ouvertes s'observe à l'échelle planétaire : au Brésil, en 1997, apparaît la Scientific Electronic Library Online (SciELO), hybride entre archive ouverte et éditeur en accès ouvert de revues. Dans les années qui suivent, ce projet s'élargit pour englober de nombreux autres pays d'Amérique du Sud, avec une approche multidisciplinaire. Cette archive ouverte est particulière, puisqu'elle est la seule, parmi les pionnières, à émerger hors d'un contexte américain ou européen, et surtout hors d'un contexte anglophone : elle témoigne de la vivacité du mouvement, encore en gestation, de l'accès ouvert. Cependant, comme on va le voir, elle n'est pas représentée dans les premières discussions autour de la naissance d'une fédération d'archives ouvertes...

On voit d'autres initiatives fleurir dans ce premier temps des archives ouvertes (comme CogPrints, archive ouverte dans le domaine de la psychologie), qui créent un réseau mondial de chercheuses et chercheurs, partageant l'information d'une manière inédite, ou plutôt, qui créent un ensemble d'initiatives ayant besoin de coordination. En effet, la disparité de ces archives est notoire, les formats sont disparates, par exemple les format TeX, créé en 1978

[14] *BibEc main page* [en ligne]. 11 décembre 1997. [Consulté le 11 janvier 2020]. Disponible à l'adresse : http://web.archive.org/web/19971211044921/http://netec.mcc.ac.uk/BibEc.html ; *WoPEc main page* [en ligne]. 11 décembre 1997. [Consulté le 11 janvier 2020]. Disponible à l'adresse : http://web.archive.org/web/19971211044714/http://netec.mcc.ac.uk/WoPEc.html.

[15] *Home Page Papers in Economics* [en ligne]. 11 décembre 1997. [Consulté le 11 janvier 2020]. Disponible à l'adresse : http://web.archive.org/web/19971211050056/http://netec.mcc.ac.uk/HoPEc.html.

[16] *Economist Jokes* [en ligne]. 11 décembre 1997. [Consulté le 11 janvier 2020]. Disponible à l'adresse : http://web.archive.org/web/19971211044513/http://netec.mcc.ac.uk/JokEc.html.

et le PDF qui apparaît au milieu des années 1990, qui coexistent avec de nombreux formats de traitement de texte. Il est complexe d'avoir accès à une liste complète des références contenues dans les archives ouvertes, ce qui pourrait intéresser de nombreuses institutions, au premier rang desquelles les bibliothèques universitaires, qui pourraient faire profiter leurs usager-e-s de ces ressources. Dans un moment où le phénomène d'informatisation des catalogues et les rétroconversions s'accentue, le problème de l'accès à ces ressources nouvelles , de leur signalement commence à se poser : comment faire profiter le plus grand nombre possible de ces *preprints* ? Le but est de transformer la communication scientifique en profitant des possibilités d'échange offertes par le numérique, et pour ce faire, de fédérer le mouvement naissant des archives ouvertes.

En effet, si l'on se place du point de vue d'un usager, les frontières sont alors assez troubles : comment trouver la référence d'un article en particulier ? Il n'existe pas de moteur de recherche centralisé, les frontières disciplinaires sont assez floues, puisque les archives ouvertes, en construction, s'ouvrent à de nouvelles disciplines, sans doute par opportunité : arXiv, au départ centrée uniquement autour de la physique, accueille rapidement des *preprints* en mathématique ou en informatique. Il faudrait donc qu'un-e usager-e connaisse parfaitement les divers endroits où trouver un *preprint*, ce qui empêche toute forme de sérendipité et bride grandement la recherche documentaire. Cela nécessite à fortiori une veille constante sur l'apparition d'autres archives ouvertes, sur l'évolution des contenus de celles qui existent déjà. Chaque personne souhaitant accéder aux contenus des archives ouvertes devait donc être, en somme, un-e professionnel-le de l'information. De la même manière, les bibliothèques ont alors été bridées dans l'accès qu'elles peuvent donner à ces contenus : hormis le fait de lister les archives ouvertes, elles ne disposent pas d'une manière unifiée de lister les ressources qui y sont contenues. Une bibliothèque universitaire spécialisée dans les sciences économiques, par exemple, ne peut guère faire mieux que d'inciter ses usagers et usagères à aller consulter RePEc.

L'archive universelle : du projet à son évolution

Pour changer cela, en juillet 1999, Paul Ginsparg, l'un des fondateurs d'arXiv, et Herbert Van de Sompel, qui travaillait au Los Alamos National Laboratory (LANL), lancent un appel à participation à une rencontre pour mettre en place une coopération entre les différentes archives de *preprints*, lançant l'éphémère Universal Preprint Service (UPS). Cette rencontre

se tient au Nouveau-Mexique, à Santa Fe, en octobre 1999[17], et rassemble des représentants d'archives ouvertes reconnues (dont arXiv, Cogprints et RePEc), des universités anglaises et américaines et d'autres institutions et associations nationales (comme la Library of Congress) et internationales, notamment l'Association of Research Libraries (ARL) ou le Scholarly Publishing and Academic Resources Coalition. Les bibliothèques sont donc partie intégrante de l'effort fédérateur des archives ouvertes, tout du moins les bibliothèques américaines et anglaises : en effet, la réunion de Santa Fe ne rassemble que des représentant-e-s d'institutions anglophones, à l'exception d'Herbert Van de Sompel, qui avait une double affiliation Université de Gand/Los Alamos National Laboratory. Alors que nous avons évoqué l'apparition de SciELO au Brésil, il est intéressant de noter l'absence de représentant-e-s d'initiatives ou institutions autre que de grandes universités américaines comme Harvard, Cornell, CalTech ou l'université de Californie, ce qui incite à questionner l'« universalisme » pourtant annoncé dans le nom de l'initiative. La réunion se donne toutefois le but de créer un standard international pour les archives ouvertes, ou plutôt, des « exigences techniques minimales pour les archives »[18].

Les débats d'alors tournent autour de problématiques qui gardent leur pertinence aujourd'hui : comment donner accès à l'information scientifique ? Comment conjuguer les besoins d'exposition des données, en l'occurrence des articles (principalement), et celui de faciliter l'accès à leur contenu ? Les débats ont fini par faire émerger un mécanisme qui sépare les acteurs du champ en deux domaines : les fournisseurs de données (*data providers*), et les fournisseurs de service (*service providers*). L'autre option envisagée était la création d'un moteur de recherche distribué, c'est-à-dire un moteur de recherche dans un réseau de pair à pair (*peer to peer* ou P2P), qui serait donc distribué entre les différentes archives ouvertes afin d'en explorer le contenu. Il convient de rappeler que dans ces années, il existe une véritable vogue des solutions en *peer to peer* pour le partage de contenus, particulièrement dans l'accès illégal ou « pirate » : grâce au réseau GNutella, des strctures comme LimeWire, Napster et d'autres sites partagent du contenu de manière décentralisée, une solution qui gagne rapidement en popularité. Les concepteurs du protocole rejettent toutefois cette solution : elle est jugée trop complexe à passer à l'échelle, puisqu'il faudrait implémenter un protocole de recherche dans toutes les archives existantes. Ce processus est

[17] *First meeting of the Universal Preprint Service Initiative* [en ligne]. [s. d.]. [Consulté le 13 janvier 2020]. Disponible à l'adresse : http://www.openarchives.org/news/ups1-press.htm.

[18] "The group agreed on minimal technical requirements for archives. These will be published separately as the "Santa Fe Conventions" and, in the next six months, will be implemented in the existing archives." *ibid.*

jugé complexe[19], et devrait s'appliquer à toutes les futures archives qui apparaissent pour rester dans le but original de permettre une interopérabilité globale entre les différentes archives ouvertes. Il y a donc une accentuation forte sur le fait que la solution proposée doit être la plus simple possible, afin de pouvoir passer à l'échelle.

La division entre *service* et *data providers* permet à chacun de rester dans sa mission effective : les *data providers,* en l'occurrence les archives ouvertes, exposent leurs données, des *preprints*, et les bibliothèques, *service providers,* viennent moissonner les données et les exposent à leur tour, fournissant ainsi une voie d'accès à l'information, et un supplément de visibilité. Cela permet de ne pas transformer un système fonctionnel et en plein essor, mais de chercher à respecter le plus possible le fonctionnement originel des archives.

Cela permet en outre aux fournisseurs de service de sélectionner les données qu'ils moissonnent, et donc de proposer un service pertinent pour leurs communautés : ainsi, une bibliothèque universitaire en sciences économiques fournirait uniquement une liste d'articles contenus dans RePEc, puisque cette archive correspondrait aux besoins des usager-e-s, alors que les contenus d'arXiv seraient moins pertinents (en dehors de ceux qui concernent l'économétrie ou l'économie quantitative). Ce système permet ainsi de ne pas stocker de l'information non pertinente, et de construire un accès adapté aux archives ouvertes, plutôt que de noyer une personne dans des contenus divers. Pour atteindre cet objectif à l'époque, trois accords sont nécessaires :

- Un accord sur le protocole, qui doit permettre de sélectionner les données à importer
- Un accord sur les critères de sélection de ces données, et sur la temporalité de leur récupération
- Un accord sur les formats des données, ou plutôt des métadonnées, qui doivent être récupérées.

Il s'agit de permettre aux personnes de choisir ce qui doit être récupéré, et aux machines de le faire correctement.

Ces discussions se terminent sur une transformation : l'Universal Preprint Service devient l'Open Archives Initiative. Ce changement de nom n'est pas neutre : il traduit, plutôt que la

[19] "The [solution] would require archives to implement a joint distributed search protocol, which is not considered to be a low-entry requirement.", *op.cit.*

création d'un service central qui permette un accès universel au contenu des archives ouvertes, un effort collaboratif pour maintenir les archives ouvertes, d'éviter des cloisonnements, notamment dus à la technologie, tout en gardant le foisonnement, la multiplication des initiatives ; il s'agit, en somme, de fédérer sans agglomérer. Il est intéressant de noter que l'américano-centrisme de la conférence de Santa Fe est d'emblée remis en question par les participant-e-s : dans les préconisations publiées, il est fait mention que la réunion suivante devrait se dérouler en Europe[20].

De cette réunion naît un cadre organisationnel : la convention de Santa Fe, en février 2000, qui est transformée en 2001 pour devenir l'Open Archives Initiative-Protocol for Metadata Harvesting, qui concrétise techniquement les trois accords, puisqu'il est le protocole qui permet de sélectionner les données. Le protocole permet de sélectionner les données à recevoir selon 6 « verbes » : GetRecord, Identify, ListIdentifiers, ListMetadataFormats, ListRecords et ListSets. Ces verbes permettent d'obtenir des enregistrements donnés (ou *records*), des informations sur l'entrepôt de données, des listes d'identifiants ou de formats de métadonnées disponibles dans l'entrepôt, ou encore des listes d'enregistrement selon certains critères (notamment de dates) ou d'ensembles (*sets*). Selon ce protocole, différents niveaux d'organisation des données sont donc possibles au sein d'un entrepôt, donc du côté des fournisseurs de données. Ces dernières peuvent être constituées en « ensembles », afin de permettre une granularité plus grande au sein, par exemple, d'une archive pluridisciplinaire. En effet, même si Cogprints est centrée autour des sciences cognitives, des fournisseurs de service pourraient n'être intéressés que par les articles concernant la psychologie, et pas les neurosciences qui sont pourtant répertoriées également sur cette archive. La logique de moissonnage, qui permet de ne moissonner qu'un fournisseur de données, est donc poussée encore plus loin puisqu'il est possible de ne moissonner qu'une partie d'un entrepôt. D'un autre côté, le poids reposant sur les archives n'est pas insurmontable, sachant que cette organisation en ensembles n'est pas obligatoire. Le protocole est, dans sa conception, une réponse au cahier des charges, qui est de fournir une exigence technique minimale. Les fournisseurs de données ont le choix de la granularité qu'ils offrent, tant qu'ils exposent. Si l'exposition respecte les logiques d'ensembles et possède des données correctement renseignées, les fournisseurs de service peuvent alors

[20]« Paul Ginsparg suggest that a next meeting should be held in Europe, in the first quarter of next year. »*op.cit.*

avoir accès de manière correcte et assez détaillée aux informations, et choisir ainsi ce qui les intéresse.

Enfin, il faut noter un aspect important d'OAI-PMH, qui a été pour beaucoup dans son succès : son asynchronie. Avant la création de ce protocole, des technologies existaient déjà pour transférer des données, et particulièrement des données bibliographiques. Le protocole Z39-50 est, dans le monde des bibliothèques, le plus connu à cet égard. Créé en 1988 par la Bibliothèque du Congrès et l'Online Computer Library Center (OCLC), le protocole permet d'échanger des données en format Machine Readable Cataloguing (MARC), de manière synchrone, c'est à dire que la requête est exécutée « en direct sur le serveur distant et les résultats sont rapatriés instantanément en retour »[21]. Ce protocole a permis au catalogage de se développer de manière collaborative, mais il a quelques problèmes : il n'est pas basé sur les technologies du web, qu'il prédate, pose des problèmes d'interopérabilité (puisqu'il n'est pas compatible avec tous les Systèmes Informatisés de Gestion de Bibliothèques, ou SIGB) et de temps d'interrogation. Ce dernier élément est particulièrement dû à son fonctionnement synchrone. Le protocole OAI-PMH, lui, fonctionne de manière asynchrone, c'est-à-dire que les données ne sont pas mises à jour en temps réel. Si un fournisseur de données met à disposition 100 notices en janvier, qu'il est moissonné le lendemain par un fournisseur de services, et qu'il en rajoute cinquante le jour suivant, le fournisseur de service gardera les 100 notices originelles, jusqu'à ce qu'il moissonne à nouveau l'entrepôt. Même si cela pose des problèmes de mise à jour des données, cette manière de fonctionner facilite le moissonnage et le dissocie de l'interrogation, qui gagne en souplesse et en rapidité. Encore une fois, la fréquence de moissonnage est laissée au fournisseur de service.

Le passage de l'idée d'un service universel à celle d'une initiative partagée se réalise donc à travers l'adoption de ce protocole commun, rapidement implémenté dans les archives ouvertes, et des critères qu'il propose pour récupérer, consulter ou lister des données. Or, un troisième accord, celui des formats de métadonnées à récupérer, est atteint par la convention de Santa Fe (et donc, par OAI-PMH qui en reprend le flambeau un an plus tard) : c'est le standard Dublin Core qui est choisi.

[21]Le protocole Z39.50. Dans : *BnF - Site institutionnel* [en ligne]. [s. d.]. [Consulté le 27 janvier 2020]. Disponible à l'adresse : https://www.bnf.fr/fr/le-protocole-z3950.

De Dublin Core comme le vaisseau du protocole

Ce standard a été développé en grande partie au sein de l'OCLC à partir de 1995 : c'est d'ailleurs de là qu'il tire son nom, Dublin ne faisant pas référence à la capital sud-irlandaise, mais à une ville des Etats-Unis, dans l'Ohio, où se trouve le siège social d'OCLC. C'est donc un standard né au sein du monde des bibliothèques, développé avec des objectifs spécifiques dans une *Request for Comments*[22] de 1998 :

> "The goals that motivate the Dublin Core effort are:
>
> - Simplicity of creation and maintenance
> - Commonly understood semantics
> - Conformance to existing and emerging standards
> - International scope and applicability
> - Extensibility
> - Interoperability among collections and indexing systems"[23]

On voit que ces objectifs (simplicité, extensibilité et interopérabilité) préfigurent ceux de la conférence de Santa Fe ; clairement, on se situe dans la volonté d'un consensus technique minimal, qui se situe dans un contexte d'émergence d'initiatives et doit donc garder une adaptabilité importante. Le but n'est pas de créer un standard de métadonnées englobant tout, précis et fermé, mais un cadre minimal qui puisse s'adapter à de nouveaux besoins, de nouvelles pratiques : c'est l'objectif d'extensibilité du Dublin Core. La simplicité et l'interopérabilité correspondent également aux objectifs du projet Open Archives Initiative : une solution facilement implémentable, et qui permette aux archives de communiquer entre elles, même avec des systèmes différents.

À l'origine, le format Dublin Core a été élaboré afin de trouver une solution à mi-chemin entre une entrée d'index et une notice bibliographique (de type Machine Readable Cataloging, ou MARC). C'est ainsi que Stuart Weibel analyse la situation dans le rapport de l'atelier de 1995 qui définit Dublin Core :

> "Une solution alternative [...] implique la création d'une notice qui est plus informative qu'une entrée d'index mais moins complète qu'une notice de catalogage. Si seul un effort humain minimal est requis pour créer de telles notices, plus d'objets pourraient être décrits, surtout si l'auteur de cette

[22]Type de document qui décrit les aspects et spécifications techniques de standards internets.

[23]WEIBEL, Stuart L., KUNZE, John A., LAGOZE, Carl, et al. *Request for Comments 2413* [en ligne]. septembre 1998. [Consulté le 15 janvier 2020]. Disponible à l'adresse : https://www.ietf.org/rfc/rfc2413.txt.

> ressource pouvait être encouragé à créer la description. Et si la description suivait un standard établi, seule la création d'une notice aurait besoin d'une intervention humaine ; des outils automatisés pourraient découvrir ces descriptions et les recueillir."[24]

Ce format intermédiaire naît d'un constat d'époque : la capacité des moteurs de recherche à retrouver des documents pose des problèmes sur un nombre de sites grandissant. Idéalement, les moteurs s'appuieraient sur une information plus riche, mais cela poserait des problèmes d'indexation, au vu de la spécificité du format (par exemple les notices MARC). Stuart Weibel soulève que l'on est en face d'un problème : l'accès de la communauté universitaire[25] à des contenus. Il conviendrait donc d'employer une forme d'information accessible à des personnes en-dehors du monde des bibliothèques. Ainsi, ces personnes pourraient créer directement des informations sur les contenus. De cette manière, la multiplication des informations et données sur le web ne pose pas d'obstacle à la recherche, puisqu'elles sont décrites dans un standard partagé. La prochaine étape est de construire un protocole pour les collecter et les rendre accessibles. À ce stade du développement du web, il est important de rappeler que certaines ressources étaient découvrables – principalement – par le bouche-à-oreille. Il est donc intéressant de voir que si le Dublin Core est un choix idéal pour atteindre les objectifs d'OAI-PMH : ce dernier est en réalité l'aboutissement, la seconde étape du projet de Dublin Core. Les deux parlent le même langage que le web, respectivement HTTP pour OAI-PMH et XML pour le Dublin Core.

Dublin Core doit servir à décrire des objets numérisés (texte, image, son) ou nativement numériques, et OAI-PMH doit permettre de retrouver ces descriptions, de permettre l'accès à ces ressources. Dans sa conception, Dublin Core est également conçu pour un usage plus large que MARC, puisque le défi de décrire les ressources va au-delà du texte imprimé : Weibel et les autres personnes travaillant sur le développement du standard avaient pris en compte la variété des ressources déjà présentes, et la possibilité que ce qui serait mis en ligne le lendemain, ou dix ans plus tard, puisse être décrit aussi bien que les ressources que nous connaissons aujourd'hui.

[24]"An alternative solution [...] involves the creation of a record that is more informative than an index entry but is less complete than a formal cataloging record. If only a small amount of human effort were required to create such records, more objects could be described, especially if the author of the resource could be encouraged to create the description. And if the description followed an established standard, only the creation of the record would require human intervention; automated tools could discover these descriptions and collect them.", traduction personnelle tirée de WEIBEL, Stuart L. *Metadata: the Foundations of Resource Description* [en ligne]. juillet 1995. [Consulté le 16 janvier 2020]. Disponible à l'adresse : http://www.dlib.org/dlib/July95/07weibel.html.

[25]Il parle dans le document des « scholarly communities ».

Ainsi, les éléments qui composent Dublin Core sont conçus pour être universels, pour être un noyau (*core*) :

> "Le nom du set d'élément, en plus d'être un reflet du patrimoine géographique, capture la notion que Dublin Core est un point départ (un noyau). Il a toujours été attendu que DC serait le noyau autour duquel les autres métadonnées croîtraient."[26]

Afin d'accomplir cet objectif, 13 éléments composent le set original du Dublin Core : titre, créateur, sujet, éditeur, contributeur, date, type, format, identifier, source, langue, relation, couverture. À ces éléments se rajoutent deux autres, un an après, en 1996 : droits et description, pour renseigner sur la propriété intellectuelle et approfondir l'élément « sujet ». Ces 15 éléments sont tous facultatifs et tous répétables, ce qui permet une souplesse remarquable : si le document que l'on décrit a été créé par 6 personnes, on peut répéter 6 fois l'élément « auteur » ; s'il n'a pas d'éditeur, il n'est pas obligatoire de le mentionner. Ainsi, le Dublin Core permet de décrire de nombreuses ressources de manière simple, à l'aide de balises Hypertext Markup Language (HTML), ce qui donnerait, pour ce mémoire par exemple :

```
<dc:title>Le protocole OAI-PMH à « l'heure du web sémantique » : bilans et perspectives</dc:title>
<dc:creator>de Lavenne de la Montoise, Vincent</dc:creator>
<dc:contributor>Svenbro, Anna</dc:contributor>
<dc:type>text</dc:type>
<dc:subject>OAI-PMH</dc:subject>
<dc:subject>Web sémantique</dc:subject>
<dc:subject>Bibliothèque numérique</dc:subject>
<dc:date>[2020]</dc:date>
<dc:language>fr</dc:language>
<dc:rights>CC BY-NC-ND</dc:rights>
```

De par sa volonté de pouvoir décrire plus que des documents textuels, le Dublin Core se place dans une volonté prospective de ses créateurs : celle de dire que le contenu de demain sera multiforme, et que les professionnel-le-s de l'information ne le géreront plus de la même manière. De fait, cet ensemble d'éléments a des applications au-delà du monde des bibliothèques et de l'information scientifique et technique, et même au-delà du monde universitaire. Il répond, dans cette fin des années 1990, au besoin d'institutions culturelles d'exposer des objets sur le web, ce qui implique, au-delà des bibliothèques, les collections des musées ou des archives. Dès 2000, la communauté muséale s'empare de ce format pour

[26] "The name of the element set, in addition to reflecting its geographic patrimony, captures the notion that DC is a starting place (a core). It has always been an expectation that DC would be a kernel around which other metadata would grow.", traduction personnelle de WEIBEL, Stuart L. Dublin Core Metadata Initiative (DCMI): A Personal History. Dans : *Encyclopedia of Library and Information Sciences* [en ligne]. 17 décembre 2009. [Consulté le 16 janvier 2020]. DOI 10.1081/E-ELIS3-120043530.

décrire les collections[27], et dans le cours des années 2000, le Dublin Core sert de format partagé entre archives, musées et bibliothèques. Nous nous situons dans une vision du web comme un lieu d'échange et de partage d'informations culturelles, qui va, de manière intéressante, plus loin que le contexte d'énonciation qui était à la fois celui de Dublin Core, puis celui d'OAI-PMH : le monde de la recherche.

Quand le protocole OAI-PMH requiert au minimum l'utilisation du Dublin Core, il faut noter que la place est laissée à d'autres formats : conscients des limites des 15 éléments choisis, les créateurs d'OAI-PMH le pensent comme un vocabulaire commun minimal, servant à l'adoption du protocole avant tout[28], de la même manière que Stuart Weibel et les créateurs de Dublin Core le voyaient comme un noyau autour duquel ajouter d'autres métadonnées. L'idée de Dublin Core, comme celle d'OAI-PMH, est de fournir, plutôt qu'un service central, une multitude d'endroits spécialisés qui peuvent donner accès à des ressources déterminées. Un moteur de recherche, ou les technologies préexistantes, trouveront difficilement un *preprint* précis en 2001. En décrivant le document en Dublin Core, et en moissonnant ces données à travers le protocole OAI-PMH dans un service particulier, un catalogue d'université par exemple, ces moteurs pourront le trouver.

Le but d'OAI-PMH est donc, à travers un protocole commun, de multiplier les bases de données de petite taille, selon les spécialités, et permettre à chaque institution de construire un service selon ses usages. Plutôt qu'une archive universelle, c'est un canevas sur lequel construire un univers d'archives communicantes, à l'image de la structure décentralisée du web.

[27]NEVILE, Liddy et LISSONNET, Sophie. Dublin Core and museum information: Metadata as cultural heritage data. *IJMSO* [en ligne]. Janvier 2006, Vol. 1, p. 198-206. DOI 10.1504/IJMSO.2006.012344.

[28]COCKERILL, Matthew, LAGOZE, Carl et SÉVIGNY, Martin. *[OAI-implementers] Reconsidering mandatory DC in OAI-PMH* [en ligne]. 5 août 2003. [Consulté le 16 janvier 2020]. Disponible à l'adresse : http://www.openarchives.org/pipermail/oai-implementers/2003-August/000953.html.

Extension du domaine d'OAI-PMH : des universités à la communauté patrimoniale

Comme nous l'avons vu, le protocole OAI-PMH, appuyé sur le format Dublin Core, est né au sein du mouvement naissant de l'accès ouvert. À la fin de l'année 2001, une réunion à Budapest, à l'initiative de l'Open Society Institute, société philanthropique, pose les bases de la « voie verte » et la « voie dorée » de l'accès ouvert (*open access*) dans la déclaration que l'on connaît sous le nom d'Initiative de Budapest pour l'Accès Ouvert[29]. Cette déclaration résume assez bien le développement de l'état d'esprit que nous avons essayé de décrire : elle fait état de la possibilité qu'offre Internet et le World Wide Web de transformer la communication entre scientifiques, ce qui était également l'objectif assumé de l'Open Archives Initiative[30]. Il n'est donc pas surprenant de retrouver, dans les deux voies à mettre en place conjointement pour développer l'accès ouvert, les standards de l'OAI :

> "Auto-archivage : D'abord, les chercheurs ont besoin des outils et d'une assistance pour déposer leurs articles de revue dans des archives ouvertes électroniques, une pratique communément connue sous le nom d'auto-archivage. Quand ces archives se conforment aux standards créés par l'Open Archives Initiative, alors les moteurs de recherche et d'autres outils peuvent traiter les différentes archives comme une seule. Les utilisateurs n'ont alors pas besoin de savoir quelles archives existent ou l'endroit où elles se situent pour utiliser leurs contenus."[31]

L'auto-archivage, ainsi décrit, est la première version de ce qui deviendra connu sous le nom de la « voie verte », complémentaire de la « voie dorée »[32]. En promouvant l'auto-archivage, on retrouve l'ambition du Dublin Core de pouvoir décrire simplement des documents, et donc de confier cette description directement aux personnes qui souhaitent mettre ces ressources en ligne. Le rôle des professionnel-le-s de l'information serait alors d'accompagner ces personnes dans la prise en main de ces outils, et dans la construction de services de découverte *via* le protocole OAI-PMH, en tant que fournisseurs de services.

---

[29]CHAN, Leslie, CUPLINSKAS, Darius, EISEN, Michael, et al. *Budapest Open Access Initiative | Read the Budapest Open Access Initiative* [en ligne]. 14 février 2002. [Consulté le 17 janvier 2020]. Disponible à l'adresse : https://www.budapestopenaccessinitiative.org/read.

[30]"The goal of the OAi is to contribute in a concrete manner to the transformation of scholarly communication", dans VAN DE SOMPEL, Herbert et LAGOZE, Carl. The Santa Fe Convention of the Open Archives Initiative. *D-Lib Magazine* [en ligne]. Février 2000, Vol. 6, nº 2. [Consulté le 13 janvier 2020]. DOI 10.1045/february2000-vandesompel-oai.

[31]"Self-Archiving: First, scholars need the tools and assistance to deposit their refereed journal articles in open electronic archives, a practice commonly called, self-archiving**.** When these archives conform to standards created by the Open Archives Initiative, then search engines and other tools can treat the separate archives as one. Users then need not know which archives exist or where they are located in order to find and make use of their contents", traduction personnelle de CHAN, Leslie, CUPLINSKAS, Darius, EISEIN Michael et al., *Budapest Open Access Initiative | Read the Budapest Open Access Initiative* [en ligne], *op.cit.*

[32]Selon la définition actuelle du consortium Couperin, "La voie dorée ou gold open access concerne des revues ou ouvrages nativement en open access, dès leur publication. Plusieurs solutions s'offrent à un éditeur ou à une revue qui souhaite s'engager dans une transition vers la diffusion en accès libre. ", *Qu'est-ce que l'OA ? – Open Access France* [en ligne]. [s. d.]. [Consulté le 31 janvier 2020]. Disponible à l'adresse : https://openaccess.couperin.org/category/open-access/.

Une archéologie numérique de la dissémination

Au moment de la déclaration de Budapest (février 2002), 5 institutions sont renseignées comme étant des fournisseuses de service[33], notamment des institutions dont certains personnels ont été impliqués dans la création du protocole (comme l'université de Southampton ou l'Old Dominion University), mais aussi une institution européenne, l'École des Études Avancées de Trieste. Un an plus tard[34], l'essaimage commence, avec d'autres institutions nord-américaines telles que l'Université du Michigan, celle de la Colombie-Britannique. En 2005, les institutions européennes participantes se sont multipliées : le Centre National de la Recherche italien, la bibliothèque universitaire de Brême sont recensées en tant que fournisseurs de service, ainsi que le Centre pour la Communication Scientifique Directe (CCSD), créé en 2000, qui a pour mission de « fournir, dans l'esprit du libre accès, des outils pour l'archivage, la diffusion et la valorisation de publications et données scientifiques. Ce dernier, devient fournisseur de service avec Open Archives en Sciences de l'Information et de la Communication (OASIC), un service qui permet d'explorer d'accéder à des articles en *preprint*, des mémoires et des thèses, puis avec Hyper Articles en Ligne (HAL), également basé sur le protocole OAI-PMH. On voit donc qu'à la suite de l'initiative de Budapest, la dissémination du protocole est rapide, et suit la logique : les différents fournisseurs de service vont moissonner dans des entrepôts précis pour construire un service qui leur correspond. Le CCSD, par exemple, avec OASIC, moissonne des bases de données françaises, alors que l'École des Études Avancées de Trieste, dont le service a été créé spécifiquement pour le département de physique, moissonne les archives ouvertes spécialisées dans le domaine, notamment arXiv. Dans un cas, l'ancrage national ou local, dans l'autre, l'ancrage disciplinaire, permettent de construire des services radicalement différents. Pour autant, l'idée d'un service universel ne disparaît pas, puisque l'université du Michigan apparaît dès 2003 avec le service OAIster, qui a pour but de rendre accessibles à la recherche toutes les ressources non découvrables par les moteurs de recherche[35], et qui moissonne, dès 2003, 142 entrepôts OAI différents.

[33] *Open Archives Initiative Service Providers* [en ligne]. février 2002. [Consulté le 17 janvier 2020]. Disponible à l'adresse : https://web.archive.org/web/20020204155940/http://www.openarchives.org:80/service/listproviders.html.

[34] *Open Archives Initiative Service Providers* [en ligne]. 17 avril 2003. [Consulté le 17 janvier 2020]. Disponible à l'adresse : https://web.archive.org/web/20030417164719/http://www.openarchives.org:80/service/listproviders.html.

[35] "Our service will **reveal digital resources previously "hidden" from users** behind web scripts (how are they hidden?). The OAI harvesting protocol we're using makes this possible", dans *OAIster Home* [en ligne]. 17 février 2003. [Consulté le 17 janvier 2020]. Disponible à l'adresse : https://web.archive.org/web/20030217011435/http://www.oaister.org/o/oaister/.

Il faut donc noter plusieurs choses :

- Même si le but des créateurs d'OAI-PMH n'était pas de créer une archive universelle, la conception du protocole rend cela possible.

- Alors qu'en 1995, les moteurs de recherche étaient presque inexistants, cela n'est plus vrai dans les années 2000.

Ce deuxième point mérite de s'y attarder. Les moteurs de recherche qui se multiplient répondent au besoin de trouver une information précise sur le web. Dans ces années, ils gagnent en popularité, mais leur conception les empêche de trouver les ressources telles que les articles, puisque ceux-ci sont contenus dans des silos, qui sont, par exemple, les catalogues de bibliothèques informatisés, ou bien les entrepôts OAI-PMH, qui ne sont pas, alors, fouillés par les moteurs de recherche. Le protocole permet donc un accès à des ressources qui, sans cela, passeraient inaperçues. Certains fournisseurs de service, plutôt que de tenter comme OAIster de « concurrencer » les moteurs de recherche, tentent de servir d'intermédiaire, comme DP9, un service de l'Old Dominion University, qui est conçu comme un *gateway service* (service passerelle) qui transforme la requête du moteur de recherche en requête compréhensible *via* le protocole OAI-PMH. Il est donc intéressant de noter que, dans les quelques années qui suivent la mise en place du protocole, le paradigme change : alors que les concepteurs du Dublin Core pensaient que les moteurs étaient incapables de trouver une information précise pour les personnes qui l'utilisent. M'amélioration de la technologie et l'usage font qu'ils deviennent les moyens privilégiés d'accéder à l'information. C'est à ce moment, en effet, que le moteur de recherche Google commence une véritable montée en puissance. Il devient donc essentiel de rester sur le chemin de cette recherche d'information, ce que l'on voit aussi bien avec OAIster qu'avec DP9.

Mais l'initiative OAIster nous renseigne également sur autre chose : la typologie des fournisseurs de services deux ans après l'introduction du protocole. Dans les 142 ressources listées en 2003, nous trouvons bien entendu le monde de la recherche en très grande majorité, mais également d'autres institutions culturelles. Ces dernières ont donc exposé leurs données selon le protocole OAI-PMH, et sont des fournisseuses de données. On retrouve ainsi de nombreuses institutions d'archives réunies au sein de projets tels qu'*Archives In London and*

the M25 area[36], mais aussi des bibliothèques numériques qui donnent accès à des contenus muséaux et archivistiques, comme la California Digital Library[37]. Des initiatives commencent donc à fédérer les professionnel-le-s de l'information dans les domaines culturels, et le contexte de création d'OAI-PMH est dépassé, encore une fois grâce à la souplesse du protocole. Au-delà de cet objectif de transformation de la communication scientifique, il permet d'améliorer la communication entre institutions, la création de projets communs. Il devient, peu à peu, le langage commun que parlent les mondes des archives, des musées et des bibliothèques. Ce sont d'ailleurs souvent des initiatives qui lient l'un, l'autre ou la totalité de ces trois professions qui émergent.

Les grandes institutions et OAI-PMH : l'exemple de la Bibliothèque nationale de France

De grandes institutions commencent également à l'employer dans ce début des années 2000. La Bibliothèque du Congrès, qui a participé à la conférence de Santa Fe, l'utilise en partenariat avec la Bibliothèque nationale de France dès 2004 pour un projet commun, le dossier France-Amérique. Formé, du côté américain, dans le cadre du projet Global Gateway, le projet fait partie des initiatives de la Bibliothèque du Congrès afin de diffuser ses contenus via le protocole OAI-PMH. Du côté de la BnF, cette première utilisation est rapidement suivie de la création de plusieurs entrepôts OAI-PMH, notamment pour la bibliothèque numérique Gallica (entrepôt OAI-NUM), mais aussi pour le catalogue général de la BnF (entrepôt OAI-CAT). Dans ce début des années 2000, l'espoir est, à travers l'utilisation du protocole, d'être mieux indexé par les moteurs de recherche, c'est à dire que les robots parcourent les entrepôts lors d'une recherche. L'utilisation du protocole donne cet espoir car, contrairement à d'autres protocoles ou standard, OAI-PMH est un standard web. En 2005, en évoquant l'utilisation du protocole, Clément Oury, alors élève conservateur des bibliothèques, explique dans un rapport de stage effectué à la bibliothèque numérique de la BnF :

> « Actuellement, les ressources disponibles dans Gallica, aussi bien que les notices du catalogue BN-Opale Plus, relèvent du «Web profond», qui n'est pas indexé parles moteurs de recherche. Néanmoins, ces moteurs, à l'instar de M.S.N., Yahoo et Google, cherchent à obtenir, grâce au protocole OAI, des références de documents indexées en Dublin Core. Ceci représente un moyen pour ces moteurs de proposer à leurs usagers des ressources de qualité – les références provenant d'un entrepôt OAI sont

[36] *AIM25 - Online research for archive collections of higher education institutions and learned societies within greater London* [en ligne]. [s. d.]. [Consulté le 17 janvier 2020]. Disponible à l'adresse : https://aim25.com/index.stm.

[37] *California Digital Library* [en ligne]. 19 février 2003. [Consulté le 17 janvier 2020]. Disponible à l'adresse : https://web.archive.org/web/20030219022653/http://www2.cdlib.org/.

> souvent bien classées dans les pages des moteurs de recherche, puisque les institutions qui mettent en place ces entrepôts présentent des gages de sérieux. »[38]

Dans ce paragraphe, on lit une véritable stratégie de présence numérique : dans un nouveau système d'accès à l'information, les institutions de savoir que sont les bibliothèques cherchent à être reconnues comme des intermédiaires de confiance par les acteurs du numérique. Pour ce faire, il est attendu que les robots des moteurs de recherche parcourent les entrepôts OAI-PMH et fassent remonter leur contenu dans les résultats. En se positionnant comme intermédiaire de confiance, la BnF met également ses documents sur le chemin des internautes. Par cette stratégie, on cherche à rendre la bibliothèque pertinente à l'heure du web, d'où l'importance d'utiliser des standards qui en sont issus.

L'adoption du standard par la BnF démontre une progression du protocole dans les institutions patrimoniales : à ce moment, c'est une technologie encore peu connue, et le fait que de grands établissements s'en emparent et collaborent à travers son utilisation fait beaucoup pour son adoption. La BnF porte également le protocole dans un autre projet collaboratif international : la construction d'une bibliothèque numérique européenne, née du souhait de Jean-Noël Jeanneney, alors président de la BnF[39]. Hormis cette dernière, le prototype d'Europeana regroupe également la Bibliothèque nationale de Hongrie et celle du Portugal[40], et est basé sur le protocole OAI-PMH. Il est présenté en 2007 à l'occasion du Salon du Livre.

Ce projet, auquel s'agrègent rapidement de nombreuses institutions à travers toute l'Europe, contribue également à répandre l'utilisation d'OAI-PMH dans la communauté patrimoniale. Ici, il convient de revenir sur ce que nous nommons la « communauté patrimoniale » : c'est un ensemble d'institutions, nationales ou non, qui sont garantes d'un patrimoine, écrit, pictural ou autre, et que nous analysons particulièrement, puisque son utilisation du protocole OAI-PMH n'est pas forcément celle de la communauté de la recherche. Bien des problématiques sont similaires, dans ces deux communautés, et leur utilisation du protocole, au départ, correspond au même but, que nous avons déjà évoqué : rendre visible des contenus sur le web. Cependant, les publics visés ne sont pas les mêmes. Van de Sompel et les autres

[38]OURY, Clément. *Le département de la bibliothèque numérique à la Bibliothèque nationale de France*. 2005, p. 70.

[39]JEANNENEY, Jean-Noël. *Quand Google défie l'Europe, par Jean-Noël Jeanneney* [en ligne]. 22 janvier 2005. [Consulté le 31 janvier 2020]. Disponible à l'adresse : https://www.lemonde.fr/archives/article/2005/01/22/quand-google-defie-l-europe-par-jean-noel-jeanneney_395266_1819218.html?xtmc=quand_google_defie&xtcr=25 et JEANNENEY, Jean-Noël. *Quand Google défie l'Europe : plaidoyer pour un sursaut*, éd. Mille et une Nuits, Paris, 2005

[40]REES, Marc. *Europeana, le prototype français de bibliothèque numérique* [en ligne]. 26 mars 2007. [Consulté le 31 janvier 2020]. Disponible à l'adresse : https://www.nextinpact.com/news/35461-europeana-bibliotheque-numerique-BNF-Google-.htm.

personnes ayant participé à la conférence de Santa Fe avaient pour objectif de transformer la communication scientifique : leur public était donc celui d'une communauté de chercheuses et de chercheurs. De grands organismes comme Gallica, ou Europeana, visent à l'inverse tous types de publics. Ces différences de public, et de contenus exposés (une image du trône de Dagobert[41] dans Gallica, un article sur les fonctions logarithmiques dans arXiv[42]), induisent des utilisations différentes du protocole et de ses composantes, notamment le Dublin Core. Ces différences d'utilisations sont bien sûr inclues dans la logique de conception du protocole, et le Dublin Core est prévu pour décrire tous types de contenus, mais à quel point cette plasticité entraîne-t-elle un appauvrissement ?

Dans l'entonnoir d'OAI-PMH

Alors que le protocole se dissémine, que les initiatives émergent et lient des communautés professionnelles, un autre phénomène se produit. La simplicité, la plasticité du protocole, qui ont garanti son succès initial, commencent à poser un problème, ou tout du moins à être identifiés comme de potentiels problèmes.

Dès 2003, Herbert Van de Sompel, un des acteurs essentiels de la création d'OAI-PMH, participe à un article dans *D-Lib Magazine*[43]qui invite la communauté à diversifier et approfondir ses usages du protocole, avec l'adjonction d'URL persistants, par exemple, ce qui permet de modifier ses données sans nuire à la cohérence des identifiants utilisés dans les bibliothèques numériques. De la même manière, Carl Lagoze, une autre figure marquante de la création du protocole, écrit dans un échange de courriels, en août 2003, que :

> "On a donné trop d'importance au lien entre Dublin Core et OAI-PMH, au détriment de l'utilité d'OAI-PMH pour disséminer des données plus riches et structurées, peut-être plus utiles."[44]

Le problème est donc l'utilisation d'OAI-PMH, en particulier l'utilisation du Dublin Core comme seul format. En le décrivant, nous avons essayé de montrer que sa simplicité était conçue comme une manière de faciliter et d'accélérer sa dissémination, mais qu'il était

[41][Trône de Dagobert]. Dans : *Gallica* [en ligne]. 0900 750. [Consulté le 31 janvier 2020]. Disponible à l'adresse : https://gallica.bnf.fr/ark:/12148/btv1b55010079x.

[42]KANAS, Stanislawa et MASIH, Vali Soltani. Solution of the logarithmic coefficients conjecture in some families of univalent functions. *arXiv:2001.11098 [math]* [en ligne]. Janvier 2020. [Consulté le 31 janvier 2020]. Disponible à l'adresse : http://arxiv.org/abs/2001.11098. ArXiv: 2001.11098.

[43]VAN DE SOMPEL, Herbert, YOUNG, Jeffrey A. et HICKEY, Thomas B. Using the OAI-PMH ... Differently. *D-Lib Magazine* [en ligne]. Juillet 2003, Vol. 9, nº 7/8. [Consulté le 10 avril 2019]. DOI 10.1045/july2003-young.

[44]"The linkage between Dublin Core and OAI-PMH has been over-emphasized at the expense of the utility of OAI-PMH for dissemination of richer,and perhaps more useful, structured data", traduction personnelle de COCKERILL, Matthew, LAGOZE, Carl et SÉVIGNY, Martin. *[OAI-implementers] Reconsidering mandatory DC in OAI-PMH* [en ligne]. 5 août 2003. [Consulté le 16 janvier 2020]. Disponible à l'adresse : http://www.openarchives.org/pipermail/oai-implementers/2003-August/000953.html.

attendu qu'il y ait un enrichissement. Or, deux ans après le lancement du protocole, force est de constater que son implémentation, dans la plupart des cas, ne dépasse pas le strict minimum. Le Dublin Core, attendu comme un noyau autour duquel construire, est utilisé à la quasi-exclusion de tout autre format de données. Il faut également que l'usage du DC soit cohérent entre fournisseur de données et de service. De cette « pauvreté » naît l'article de Van de Sompel, qui tente de proposer des enrichissements, des détournements de l'utilisation du protocole. Les échanges de courriels incluant Carl Lagoze, eux, mentionnent la possibilité de cesser, dans le protocole, de rendre le Dublin Core obligatoire, pour encourager d'autres formats plus riches, plus structurés, qui ne soient pas sémantiquement plats. Il importe ici de préciser les problèmes que peut causer le format, après avoir discuté ses avantages. La souplesse de l'ensemble d'éléments lui permet théoriquement de décrire toute ressource, qu'elle soit numérique, numérisée, physique, mais chacune de ces formes a des spécificités qui ne sont pas prises en compte par les 15 éléments répétables. L'élément *dc:creator*, par exemple, ne sera pas interprété de la même manière dans un film, un morceau de musique ou un livre. Chaque secteur aura ses conventions ; on suppose qu'en cinéma, le créateur est la personne qui réalise le film, alors que pour un livre le créateur est l'écrivain-e de l'ouvrage. Mais que faire quand il y a plusieurs personnes responsables d'une création ? En cinéma, comment qualifier les acteurs et actrices ? Chaque choix est potentiellement légitime, mais il est également laissé à la libre interprétation des personnes qui décrivent la ressource, ce qui affecte grandement la cohérence. De plus, dans le cas d'une création multiple, il n'est pas possible de donner une qualité différente : soit on est *dc:creator*, soit on est *dc:contributor*. Cela implique parfois de renoncer à créer une information fine, d'autres fois de répéter de nombreuses fois un champ. Au moment de la conception du Dublin Core, le fait de ne pas avoir la précision d'un format comme MARC était un avantage, puisque cela le rendait plus simple d'utilisation. Mais cette simplicité était supposée garantir une description simple de ressources avant enrichissement. En somme, elle permettait à chacun-e de rendre visibles des documents qui seraient autrement restés confidentiels. Mais il appartenait ensuite à d'autres personnes, et notamment les professionnel-le-s de l'information, de se saisir de ces descriptions pour les enrichir. En somme, il fallait planter le noyau du Dublin Core pour ensuite faire pousser autour une véritable arborescence.

À la place, l'adoption massive du format, et l'exclusion d'enrichissements, a constitué un appauvrissement des données exposées sur le web. La réflexion entamée par Lagoze, Cockerill et Sévigny en 2003 n'aboutit pas, et le Dublin Core simple reste une brique

fondatrice du protocole OAI-PMH, et garde son statut obligatoire (*mandatory*). En 2004, une étude[45] confirme les craintes des créateurs et de la communauté concernant l'utilisation du Dublin Core : plus de la moitié des fournisseurs de données n'allaient pas au-delà de l'utilisation de deux éléments : *dc:creator* et *dc:identifier*. Il y a donc un véritable problème de pauvreté des données mises en place. Au-delà de ces données extrêmement pauvres, la grande majorité (74%) des fournisseurs ne dépassaient pas l'utilisation de 5 éléments (sur 15) qui sont *dc:creator*, *dc:identifier, dc:title, dc:date* et *dc:type*. Le problème réside donc dans une utilisation à minima du Dublin Core. Ce n'est donc pas le format minimal, mais souvent le seul format dans les entrepôts OAI-PMH. Dublin Core, sachant que celui-ci, malgré sa simplicité, permet une description plus poussée que ce que transmettent, en 2004, les fournisseurs de données. Les champs *dc:subject*, *dc:language* ou encore *dc:rights*, qui peuvent permettre une meilleure description des contenus, ne sont que marginalement utilisés par la communauté. Alors que l'équipe créatrice du Dublin Core avait insisté pour ajouter deux éléments pour la gestion de la description et des droits, ceux-ci ne sont tout simplement pas utilisés dans la mise à disposition des données pour leur moissonnage. La question se pose alors : quel est le bénéfice de fournir une information parcellaire, plate, à disposition ? N'y a-t-il pas une possibilité d'augmenter la qualité des données mises en ligne ?

[45]WARD, Jewel. Unqualified Dublin Core usage in OAI-PMH data providers. *OCLC Systems & Services: International digital library perspectives* [en ligne]. Mars 2004. [Consulté le 24 juillet 2019]. DOI 10.1108/10650750410527322. World.

D'OAI-PMH AU WEB SEMANTIQUE : DES EVOLUTIONS CONTRARIEES ?

Avant de considérer les possibilités qu'offre le web sémantique, à la fois dans les usages généraux du web et dans les milieux des professionnel-le-s de l'information, il convient de faire une sorte de point d'étape sur les manières dont l'utilisation du protocole OAI-PMH a pu évoluer, jusqu'à aujourd'hui, afin de comprendre ses avantages et ses limites en 2020, près de vingt ans après sa conception.

QUEL(S) USAGE(S) D'OAI-PMH, DE LA RECHERCHE AU PATRIMOINE ?

Open Access et archives institutionnelles : vers un repli des données ?

Pendant la période d'essaimage du protocole OAI-PMH, on assiste à la création d'archives ouvertes de petite taille, souvent à l'échelle d'un établissement. Nous avons vu quelques exemples parmi les précurseurs, par exemple HAL ou le *Consiglio Nazionale delle Ricerche*[46] italien, qui dans le début des années 2000 ont un nombre relativement restreint de documents. En France, un mouvement semblable se développe dans les années 2000. En 2004, l'INSA de Lyon ouvre SYNAPSE, son archive institutionnelle locale. Dans le même temps, des réseaux d'archives se créent, notamment au travers des universités numériques thématiques (UNT), créées à partir de 2003 sous l'égide du Ministère de l'Enseignement Supérieur et de la Recherche[47], et qui se regroupent pour former un réseau de portails, entre les différentes universités numériques et leurs adhérents, avec des échanges de métadonnées via le protocole OAI-PMH[48] en 2005-2006. Deux ans plus tard, les initiatives de l'INSA, des UNT et d'autres acteurs comme l'Université de Valenciennes et Rennes 1 se rejoignent pour

[46] Qui documente l'implémentation d'OAI-PMH sur : CONSIGLIO NAZIONALE DELLE RICHERCHE. *ExtGASoai - OAI-PMH gateway per ExtGAS: organizzazione logica del software* [en ligne]. 2007. [Consulté le 27 février 2020]. Disponible à l'adresse : http://www.cnr.it/prodotto/i/160824.

[47] À ce sujet, voir ISAAC, Henri. *L'université numérique*. Rapport à Madame Valérie Pécresse, ministre de l'Enseignement Supérieur et de la Recherche. [S. l.] : [s. n.], 2008.

[48] *Historique du projet ORI-OAI | ORI-OAI : Valoriser le patrimoine numérique scientifique, pédagogique et documentaire des universités par un réseau de portails communicants* [en ligne]. 7 septembre 2011. [Consulté le 6 février 2020]. Disponible à l'adresse : Historique.html.

créer l'Outil de Référencement et d'Indexation de portails OAI-PMH (ORI-OAI), qui peut fonctionner en corrélation avec des systèmes de type Nuxeo ou Alfresco. Il a été créé avec les objectifs suivants :

> L'Outil de Référencement et d'Indexation pour un réseau de portails OAI-PMH **ORI-OAI** vise la mise en place d'un système ouvert, en logiciel libre, permettant :
>
> - de **gérer** et de **publier** tous les documents numériques produits par les établissements universitaires,
> - de les **partager** avec d'autres établissements,
> - de les **valoriser** par une indexation de qualité,
> - de les **rendre accessibles**, à distance et selon les droits définis, dans des interfaces ergonomiques.[49]

L'idée est donc de favoriser la création d'archives ouvertes institutionnelles, et de créer un réseau pour les fédérer, et en même temps participer à améliorer la qualité des métadonnées échangées, en s'appuyant notamment, en plus du Dublin Core simple, sur deux standards de métadonnées : la norme LOMFR[50] et la recommandation TEF[51]. Le premier est la version française du format *Learning Object Metadata*, et donc prévu pour décrire des ressources pédagogiques ; le second, Thèses Électroniques Françaises, est une modélisation développée par l'ABES pour décrire et valoriser les thèses dans le contexte académique français. C'est donc un projet qui prévoit d'utiliser le protocole OAI-PMH d'une manière plus riche que la plupart des implémentations qui s'en tiennent à quelques éléments du Dublin Core. Dans le contexte des universités, c'est également une manière de valoriser des documents spécifiques à ce contexte, à savoir des ressources particulières qu'on ne trouve pas, sauf cas spéciaux, dans des bibliothèques nationales ou patrimoniales, par exemple. L'exposition numérique de ces contenus participe du projet initial d'OAI-PMH, mais en respectant la logique d'enrichissement qui était initialement contenue dans le protocole, sachant que l'idée d'un réseau de portails partageant des contenus similaires rend intéressant cet enrichissement en termes de métadonnées : un degré de précision dans les jurys de soutenance, les sujets abordés par un contenu pédagogique, sont clairement intéressants pour des événements qui voudraient donner accès à ce contenu à des usager-e-s. Le dernier point, qui porte sur les droits, va également dans le sens d'une implémentation maximale du protocole, avec l'utilisation du champ <dc:rights>, qui est, comme on l'a vu, très peu utilisé.

---

[49] *ORI-OAI | Université Numérique Ingénierie et Technologie* [en ligne]. 5 décembre 2010. [Consulté le 6 février 2020]. Disponible à l'adresse : https://web.archive.org/web/20101205040037/http://www.unit.eu/fr/enseignant/ori-oai.

[50] Voir la présentation donnée sur le site EduScol : LOMFR (Learning Object Metadata). Dans : *éduscol, le site des professionnels de l'éducation* [en ligne]. [s. d.]. [Consulté le 14 février 2020]. Disponible à l'adresse : https://eduscol.education.fr/numerique/dossier/archives/metadata/ressources-educatives-numeriques/lomfr-learning-object-metadata.

[51] Voir la présentation sur le site de l'ABES : *TEF : Métadonnées des thèses françaises* [en ligne]. [s. d.]. [Consulté le 14 février 2020]. Disponible à l'adresse : http://www.abes.fr/abes/documents/tef/.

ORI-OAI est donc une alternative à la création d'un portail dans HAL, c'est à dire un espace thématisé, regroupant toutes les contributions d'un établissement ou laboratoire particulier, mais au sein de l'archive ouverte. Ce choix est offert de manière quasiment simultanée avec la création d'Hyper Articles en Ligne, puisque dès 2003, l'Institut National de Physique Nucléaire et de Physique des Particules (IN2P3) dispose d'un portail sur HAL, suivi par l'Institut national de recherche en informatique et automatique (INRIA) l'année suivante, mais aussi par l'Université Jean Monnet, l'École Nationale Supérieure de Lyon en 2006, et encore d'autres établissements[52].

Ces deux choix reflètent des logiques différentes. Dans un cas, privilégier la décentralisation de l'information : à chaque université, une archive institutionnelle pour valoriser ses productions, qui participent au rayonnement de l'université. Dans l'autre, confier des ressources (notices ou documents en plein texte) à une structure centralisée, à faire gérer non pas au sein de l'établissement, mais dans une institution distincte, le CCSD. Dans le premier cas, on est véritablement dans la philosophie de création du protocole, dans le second, l'utilisation d'OAI-PMH est limitée à un moissonnage des métadonnées de documents d'universités par le CCSD pour les afficher en tant que notices dans HAL, ce qui reste un pis-aller par rapport au fait d'avoir le plein texte disponible dans l'archive ouverte. Comme à la BnF, le protocole OAI-PMH est aussi, dans HAL, un outil de gestion interne des données : la séparation des données de l'archive en plusieurs entrepôts (ou en plusieurs sets dans le même entrepôt) permet de créer les différents portails et collections des universités, institutions ou laboratoires, qui en retour peuvent moissonner ces métadonnées pour les verser dans un outil de découverte qui renvoie vers HAL.

Au cours de son existence, ORI-OAI est utilisé, en dehors des institutions fondatrices, par plusieurs établissements, parmi lesquels l'Université de Lorraine, alors constituée en tant que Pôle de Recherche et d'Enseignement Supérieur (PRES), l'université Lille 2, Panthéon-Assas mais aussi l'université de Strasbourg. Dans ces mises en application , souvent, les thèses sont regroupées avec d'autres travaux universitaires (mémoires, habilitations à diriger des recherches...) produites dans le cadre d'une institution. On est donc bien dans le cas d'une valorisation, par les établissements, de leur production, d'une manière décentralisée. Dans ce cadre, le protocole OAI-PMH sert à transporter des notices à afficher sur une interface, et depuis laquelle on peut avoir accès au document. On peut voir que ces notices

[52] Voir la liste des portails, avec leur date de création, dans : *Archive ouverte HAL - Les portails de l'archive* [en ligne]. [s. d.]. [Consulté le 27 février 2020]. Disponible à l'adresse : https://hal.archives-ouvertes.fr/browse/portal.

répondent bien au projet original d'enrichir les métadonnées, par exemple avec ce *record* extrait de THEOREME[53] (l'archive ouverte de l'Université Polytechnique des Hauts-de-France, anciennement université de Valenciennes).

```
<tef:thesisRecord>
  <dc:title xml:lang="fr">Anseÿs de Gascogne : Édition critique et étude de la seconde partie de la mise en prose copiée par David Aubert (à partir du f°320 r°), d'après le manuscrit 9 (Bruxelles, KBR)</dc:title>
  <dcterms:alternative xml:lang="en">Anseÿs de Gascogne : David Aubert's copy in prose : a critical edition and study of the second part of the tale (from f°320 r°) based on manuscript 9 (Bruxelles, KBR) </dcterms:alternative>
  <dc:subject xml:lang="fr">Geste des Lorrains</dc:subject>
  <dc:subject xml:lang="fr">David Aubert</dc:subject>
  <dc:subject xml:lang="fr">Mise en prose</dc:subject>
  <dc:subject xml:lang="fr">Littérature épique</dc:subject>
  <dc:subject xml:lang="en">Lorraine cycle</dc:subject>
  <dc:subject xml:lang="en">David Aubert</dc:subject>
  <dc:subject xml:lang="en">Copy in prose</dc:subject>
  <dc:subject xml:lang="en">Epic litterature</dc:subject>
- <tef:sujetRameau>
  - <tef:vedetteRameauTitre>
      <tef:elementdEntree autoriteExterne="029207800" autoriteSource="Sudoc">Geste des Lorrains</tef:elementdEntree>
      <tef:subdivision autoriteExterne="02779038X" autoriteSource="Sudoc" type="SubdivisionDeSujet">Histoire et critique</tef:subdivision>
      <tef:subdivision autoriteExterne="027253139" autoriteSource="Sudoc" type="subdivisionDeForme">Thèses et écrits
        académiques</tef:subdivision>
    </tef:vedetteRameauTitre>
  - <tef:vedetteRameauPersonne>
      <tef:elementdEntree autoriteExterne="029235235" autoriteSource="Sudoc">Aubert, David </tef:elementdEntree>
      <tef:subdivision autoriteExterne="027253139" autoriteSource="Sudoc" type="autrePartieDuNom">Thèses et écrits
        académiques</tef:subdivision>
    </tef:vedetteRameauPersonne>
  </tef:sujetRameau>
  <dcterms:abstract xml:lang="fr">Avec ses quatre témoins, le récit épique d'Anseÿs de Gascogne est l'un des succès du Moyen-Âge. C'est une conclusion du Cycle des Lorrains nettement en faveur des comtes du Nord, et Philippe le Bon en a naturellement commandé une mise en prose. La copie de David Aubert datée de 1465 nous est parvenue dans un luxueux manuscrit conservé à Bruxelles (KBR, 9), quatrième volume des Histoires de Charles Martel. La seconde partie du récit – à partir de la bataille en Santerre – sur laquelle s'appuie notre étude, ne peut être considérée comme un simple dérimage. En effet, elle révèle les goûts de la société bourguignonne du XVe siècle, s'organise en chapitres et modernise la langue, le vocabulaire et surtout les techniques littéraires. Nous sommes donc en présence d'un récit davantage ancré dans l'époque de la copie que dans celle de son modèle en vers. La présente édition réunit une description du manuscrit, une étude littéraire commentant la filiation de la prose et du vers, une analyse linguistique ainsi qu'un glossaire, un index des noms propres et une table des proverbes.</dcterms:abstract>
```

Partie d'un *record* OAI-PMH exposé dans THEOREME

On voit ici les champs Dublin Core que nous avons déjà étudiés, dont *title*, *subject*, *abstract*, mais on remarque également le format TEF, qui permet notamment de rajouter des sujets du Répertoire d'autorité-matière encyclopédique et alphabétique unifié (RAMEAU), et donc de fournir une meilleure indexation des ressources décrites. On voit donc que la réalisation du projet a réussi à concrétiser ses objectifs en termes de description et d'indexation. Mais la structure décentralisée proposée par ORI-OAI n'a pas eu une existence extrêmement pérenne, puisque de nombreuses universités qui avaient utilisé cette solution ont, en 2020, cessé de l'entretenir, voire l'ont fermée définitivement. Cette solution persiste, dans certains établissements, pour mettre en ligne des documents de littérature grise, comme les mémoires ou les thèses d'exercice, comme la plateforme AURORE de l'Université de Limoges[54]. Mais chez les grands établissements qui avaient fait le choix de cette solution, on se tourne vers

[53] *THEOREME* [en ligne]. [s. d.]. [Consulté le 27 février 2020]. Disponible à l'adresse : https://theoreme.uphf.fr/index.html.

[54] *Aurore* [en ligne]. [s. d.]. [Consulté le 27 février 2020]. Disponible à l'adresse : http://aurore.unilim.fr/ori-oai-search/index.html.

une situation centralisée : l'Université de Lorraine, par exemple, cesse d'alimenter PETALE et fait le choix de HAL comme archive ouverte institutionnelle en 2018[55].

Ce choix semble correspondre à une recentralisation, l'une des deux stratégies que nous évoquions plus haut. Cette recentralisation a du sens, dans la mesure où HAL est une plateforme centrale, bénéficiant de financements et d'une attention particulière dans le cadre des différent-e-s plans, programmes et lois pour l'accès ouvert. Dans ce cadre, HAL travaille son référencement, et peut donc améliorer la visibilité des ressources versées, ce qui est beaucoup moins facile pour une université. De la même manière, les ressources qui sont versées dans cette archive ouverte centrale bénéficient des solutions de conservation pérenne qui sont offerts par le Centre Informatique de l'Enseignement Supérieur (CINES)[56]. Ce dernier point n'est pas anodin, et est central dans l'évolution de la manière d'utiliser OAI-PMH. L'inflation du nombre de contenus numériques augmente les coûts de maintien d'une structure pour les collecter, les conserver, et la conservation pérenne de documents numériques pose de nombreux défis techniques qu'il peut être compliqué d'assumer dans un établissement de taille moyenne, ou même dans un grand établissement. OAI-PMH n'est pas un protocole qui conserve des données numériques, mais il a été créé dans un but d'échange de métadonnées, à un moment où l'inflation du nombre de contenus était à ses débuts : certains des premiers exemples de fournisseurs de services ne regroupaient que quelques centaines, tout au plus quelques milliers de notices. Les universités, particulièrement dans la logique des fusions, se retrouvent à gérer des corpus extrêmement importants, et l'enjeu de visibilité de l'université est résolu par le portail, c'est-à-dire la couche graphique que propose HAL, qui reprend le visuel de chaque institution, et permet de regrouper les contributions particulières de l'université, avec des possibilités de curation de ce contenu (collections, divisions par labos...). Pour faciliter l'accès à la recherche, la contribution à une structure centrale fait également sens : pour les chercheurs et chercheuses, l'utilisation d'un outil unique est une plus-value. La simplification du parcours utilisateur est un processus-clé dans l'objectif d'avancer vers l'accès ouvert avec une approche incitative pour les chercheurs. Et dans la logique de la voie verte, celle de l'auto-archivage, la visibilité de l'archive ouverte est cruciale, et il n'est pas possible à une multiplicité d'archives institutionnelles de

[55]« Attention : Pétale n'est plus alimenté depuis mars 2018. En effet l'Université de Lorraine a fait le choix de HAL comme archive institutionnelle. », retrouvé sur *PETALE* [en ligne]. 26 décembre 2018. [Consulté le 6 février 2020]. Disponible à l'adresse : https://web.archive.org/web/20181226203412/http://petale.univ-lorraine.fr/index.html.

[56]Décrites dans *Archivage pérenne | CINES* [en ligne]. [s. d.]. [Consulté le 6 février 2020]. Disponible à l'adresse : https://www.cines.fr/archivage-perenne/.

développer une force de frappe comparable. Dans la logique d'un point d'accès unique, ou préféré, il semble opportun de réunir des contenus dans une plateforme centrale, vers laquelle on dirige les chercheurs, chercheuses et étudiant-e-s pour inciter à la découverte de ces contenus, et ainsi la présenter comme une alternative possible à l'accès payant. De la même manière, dans un objectif de transformation de l'évaluation par la mise en place de critères sur l'accès ouvert, il est compliqué de gérer tout un réseau d'archives institutionnelles décentralisées, qui viendraient s'ajouter à une plateforme nationale (HAL), mais aussi à des archives ouvertes thématiques et disciplinaires (RePEc, arXiv...) qui sont déjà identifiées comme centrales. Il est plus simple pour la plateforme de tisser des partenariats directement avec ces plateformes, ce qui est par exemple mis en place au sein de HAL, notamment avec arXiv[57]. Dans un contexte où le référencement, la centralité et l'utilisation massive d'une plateforme en font un acteur important de l'accès ouvert (lui-même reconnu comme un enjeu clé pour les bibliothèques universitaires), on comprend que la solution ORI-OAI ne répond pas de manière optimale à ces besoins.

Un changement de paradigme pour le référencement

On peut d'ailleurs questionner, à ce sujet, la pertinence même, non d'ORI-OAI, mais d'OAI-PMH lui-même. Conçu dans un web profondément décentralisé, autour de la notion d'un service pertinent, de la visibilité de contenus de la recherche, il représente aussi la notion que les ressources seraient plus visibles par les moteurs de recherche. En vérité, trois ans après la version définitive du protocole, Google, qui était déjà le principal outil dans ce domaine, crée son protocole de découverte, Sitemaps[58], et ne *crawle* donc pas le contenu des entrepôts OAI-PMH pour les faire remonter dans ses résultats. L'espoir d'améliorer le référencement à travers l'utilisation du protocole ne se concrétise pas, d'autant plus que les autres moteurs de recherche, notamment Yahoo et Microsoft, adoptent également cette technologie afin de faire remonter des résultats plus pertinents[59]. Afin d'obtenir un meilleur référencement, il faut donc se conformer à ce protocole.

[57]Voir la documentation attenante, *Transférer le dépôt vers arXiv – HAL Documentation* [en ligne]. [s. d.]. [Consulté le 6 février 2020]. Disponible à l'adresse : https://doc.archives-ouvertes.fr/deposer/transfert-hal-arxiv/.

[58]Voir l'annonce de la création du protocole en 2005, *Google Blog: Webmaster-friendly* [en ligne]. 8 juin 2005. [Consulté le 6 février 2020]. Disponible à l'adresse : https://web.archive.org/web/20050608015054/http://googleblog.blogspot.com/2005/06/webmaster-friendly.html.

[59]*Major Search Engines Unite to Support a Common Mechanism for Website Submission – News announcements – News from Google – Google* [en ligne]. [s. d.]. [Consulté le 6 février 2020]. Disponible à l'adresse : https://googlepress.blogspot.com/2006/11/major-search-engines-unite-to-support_16.html.

La logique d'OAI-PMH était aussi de construire un service véritablement adapté à une communauté d'usagers : une recherche sans « bruit ». Cela impliquait des points d'accès extrêmement spécialisés, depuis lesquels l'usager-e trouverait des contenus pertinents. Mais, à l'heure actuelle, le moteur de recherche généraliste ou spécialisé est souvent l'accès direct, et unique, des chercheurs à des contenus. Le fait de créer un service personnalisé, disponible sur le site d'une université, par exemple, se situe dans une logique de points d'accès multiples qui n'est plus pertinente à l'heure où le web se resserre autour de quelques sites qui sont les passerelles vers le reste du web. Ainsi, même si la pertinence d'un service construit avec des données issues d'entrepôts OAI-PMH est totale pour une communauté, il conviendrait d'en faire la publicité auprès de celle-ci, afin que son usage, et donc la peine que l'on prend à construire ce service, reste pertinent. Si l'on abandonne la logique des points d'accès multiples, et qu'on se positionne autour de plateformes centrales comme HAL, le fait d'utiliser OAI-PMH pour transférer des notices de travaux de manière massive, l'intérêt est, là aussi, limité : en effet, il est beaucoup plus utile d'avoir accès au document dans son intégralité, en plein texte, qu'à une notice qui peut éventuellement renvoyer, à travers un URL, vers le document lui-même. La logique de recentralisation, on l'a vu, est également guidée par une logique de réduction des coûts : centraliser les documents permet de faire porter les coûts d'hébergement, d'archivage, de mise à disposition à une seule structure. La concentration des articles et travaux de recherche permet aussi de rendre de nombreux services aux chercheurs et chercheuses, notamment de fournir une liste de publications qui serve de curriculum vitae, relier ces publications à un ou plusieurs identifiants pérennes (ORCID, IdHAL, IdRef...). Ces services sont plus complexes à rendre si les travaux sont éparpillés entre plusieurs bases de données, où il faudrait par exemple individuellement les rattacher à des identifiants.

Les limites des archives institutionnelles

Dans le paysage actuel de la recherche, de la science ouverte, du numérique, nous sommes invités à questionner ce qu'OAI-PMH a contribué à fonder : les archives institutionnelles (*institutional repositories*). En 2016, Eric Van de Velde, qui participait à la conférence de Santa Fe en 1999 en tant que représentant de CalTech, publie un billet de blog[60] intitulé « Let IR RIP » (laissez les archives institutionnelles reposer en paix), qui liste un ensemble de

[60]VAN DE VELDE, Eric. SciTechSociety: Let IR RIP. Dans : *SciTechSociety* [en ligne]. 24 juillet 2016. [Consulté le 24 juillet 2019]. Disponible à l'adresse : http://scitechsociety.blogspot.com/2016/07/let-ir-rip.html.

raisons pour lesquelles ce type d'archives doit disparaître : manque d'enthousiasme, problèmes de gestion et d'ergonomie, faible utilisation, coûts trop élevés, problèmes de contrôle des droits, et enfin manque d'interactions. Ce dernier problème est intéressant à plus d'un titre : on a vu dès la création des archives ouvertes une composante sociale, à travers la création de JokEc par exemple[61]. L'évolution du web vers une forme plus sociale a fait émerger des réseaux sociaux dédiés à la recherche, qui sont devenus une nouvelle manière de réaliser l'objectif défini à la conférence de Santa Fe, c'est-à-dire de transformer la communication scientifique. La manière dont OAI-PMH a tenté de répondre à ce défi n'est pas la création d'un réseau social ; il est donc normal que les archives institutionnelles ne soient pas que cela.

Mais les problématiques de gestion des droits, d'utilisabilité, tout cela prend – en partie – sa source dans la mise en place du protocole, dans son implémentation minimale, qui ne définit pas finement ses contenus, qui ne transporte que des métadonnées et pas des ressources. Le manque d'enthousiasme, la faible utilisation découlent, comme on l'a vu, d'un changement des pratiques de recherche, conséquences de la logique d'un point d'accès unique. Van de Velde soutient que des grandes archives ouvertes, disciplinaires ou nationales, sont bien plus efficaces qu'une fédération d'archives institutionnelles. ORI-OAI correspond à ce modèle de fédération, dont on a étudié les limites plus tôt. Pousser à ce système peut également prendre une forme obligatoire : par l'adoption d'un mandat de dépôt, on peut imposer aux chercheurs et chercheuses de verser des publications dans une archive institutionnelle, mais l'obligation ne produit pas d' « effet de réseau ». On entend par ce terme le fait qu'une personne qui utilise le service en devienne un-e ambassadeur-rice, ce qui arrive rarement si on adopte un système où le dépôt est une injonction, et non une incitation.

Le but visé étant de créer un engouement autour de la science ouverte, il paraît plus logique d'encourager les chercheurs et chercheuses que de les forcer à utiliser des outils qui sont parfois peu intuitifs, tant dans le dépôt que dans la recherche. Pour changer cet état de fait, on peut investir lourdement dans les archives institutionnelles, et ainsi tenter de créer des plateformes et des manières de rechercher du contenu qui soient intuitives et ergonomiques : mais vient alors la barrière du coût qu'évoque Van de Velde, à conjuguer avec l'offre disponible : le secteur qui fournit ces outils étant extrêmement concentré, il n'est pas simple de trouver une solution satisfaisante à un coût raisonné ; développer en interne requiert des

[61] Voir page 13 de ce mémoire.

compétences, et représente tout de même un certain coût. Si les outils que sont les archives institutionnelles se limitent à une portion congrue, il devient alors légitime de questionner leur pertinence.

Au-delà des problématiques qu'évoque Van de Velde, on note également le positionnement, en 2016, de l'un des fondateurs du protocole OAI-PMH sur l'héritage qu'il laisse, dans la communauté universitaire tout au moins. L'objectif de créer un univers d'archives qui se répondent est, à l'heure actuelle, fortement critiqué dans la profession, et celui d'augmenter la visibilité des documents dans un univers numérique est remis en cause à la fois par les évolutions de la technologie et ceux de l'usage. Mais il convient également de voir la manière dont le protocole s'utilise aujourd'hui dans la communauté patrimoniale, où il avait également essaimé.

Quelle utilisation ?

Il est intéressant de prendre un exemple actuel pour comprendre la manière dont, après 20 ans, ce protocole est utilisé dans les établissements. L'exemple de la Bibliothèque Interuniversitaire de Santé, maintenant partie intégrante de l'Université de Paris, est intéressant à ce titre.

Dans le cadre de son service Histoire de la Santé, la bibliothèque cherche à valoriser l'exposition des données en ligne, et noue des partenariats avec des institutions culturelles à cette fin, afin de permettre aux collections patrimoniales d'être vues par le plus grand nombre. Dans ce but, la bibliothèque s'est dotée d'un moissonneur et de deux entrepôts OAI-PMH, correspondant respectivement à la banque d'images de la bibliothèque, et à la bibliothèque numérique Medica[62]. Si nous reprenons les termes du protocole, la BIUS est à la fois fournisseuse de service et de données. Les deux entrepôts exposent les données de la bibliothèque, mais Medica fonctionne également en moissonnant des contenus d'autres sources, particulièrement de bibliothèques qui contribuent à la Medical Heritage Library. Cela permet de fournir un service à valeur ajoutée pour les utilisatrices et utilisateurs de Medica, qui ne sont pas obligatoirement au courant des autres endroits où des contenus pertinents peuvent être trouvés.

[62] *Bibliothèque numérique Medica — BIU Santé, Paris* [en ligne]. [s. d.]. [Consulté le 10 février 2020]. Disponible à l'adresse : https://www.biusante.parisdescartes.fr/histoire/medica/index.php.

Il est très intéressant de se pencher sur les détails du protocole à la BIUS :

- Les entrepôts sont constitués de références en Dublin Core simple, ce qui est très représentatif de l'implémentation du protocole de manière générale

- La bibliothèque a construit un mapping spécifique à chaque entrepôt moissonné. En effet, les utilisations du Dublin Core changent selon les fournisseurs de données, et, pour que les notices apparaissent correctement dans la bibliothèque numérique, il est nécessaire de réaligner les champs.

- La conception du moissonneur fait qu'à chaque passage, il récupère la totalité des notices contenues dans un entrepôt. Il faut donc faire un tri après le moissonnage afin de n'avoir que les notices les plus récemment ajoutées, avant que des agents humains décident, parmi ce nombre, ce qui rejoindra Medica.

En plus de la valeur ajoutée du service, enrichi de contenus extérieurs, le moissonnage des données de la BIUS est intéressant par le passage qu'il provoque sur le site web : à un moment, il a été mesuré que 8% du trafic du site provenait de Gallica[63].

L'effet de décentralisation/recentralisation que nous avons observé précédemment est à l'effet ici aussi. Les acteurs avec lesquels est en contact la BIUS sont de grandes institutions, notamment Gallica et Internet Archive, où sont conservés les contenus de la Medical Heritage Library.

Ces grandes institutions permettent une visibilité bien plus forte des contenus, particulièrement via leur référencement dans les moteurs de recherche. Cela questionne la sorte de décentralisation que proposait OAI-PMH ; en effet, le protocole constituait une recentralisation de métadonnées diverses afin de fournir du service, mais son utilisation a débouché sur une multiplicité d'initiatives, la création de nombreuses bibliothèques numériques, portails ou de petites archives ouvertes. Mais à l'heure actuelle, la tendance n'est pas au développement d'initiatives de petite taille, mais bien plutôt à la recentralisation des contenus. En faisant des recherches sur la dissémination du protocole, il a été intéressant de constater que la plupart des initiatives lancées via OAI-PMH entre 2003 et 2005, répertoriées sur Open Archives Initiative, ont aujourd'hui disparu. L'agrégateur OAIster, que

[63]GHUZEL, Olivier. *Entretien sur l'utilisation d'OAI-PMH au sein de la Bibliothèque Interuniversitaire de Santé*. 31 janvier 2020.

nous avons étudié, a été pris en main par OCLC, et intégré à l'environnement général de ses produits. On a également noté le primat actuel de HAL sur les archives ouvertes propres aux universités. De la même manière, dans la communauté du patrimoine, de nombreux entrepôts OAI-PMH existaient, principalement pour être moissonnés par la BnF dans le cadre de la bibliothèque numérique Gallica : les bibliothèques municipales de Dijon, de Lyon, la Bibliothèque Francophone Multimédia (BFM) de Limoges[64] sont toutes, dans le vocabulaire du protocole, des fournisseuses de données. Mais de la même manière qu'ORI-OAI, la solution décentralisée présente des questions de coût, de maintenance des serveurs, et cette logique de bibliothèques numériques multiples est aujourd'hui en question. HAL est la structure centralisée la plus à même de répondre aux besoins des communautés universitaires, et, dans l'écosystème des bibliothèques patrimoniales, la BnF est le partenaire naturel. La solution Marque Blanche[65], créée en 2013 avec Numistral (l'émanation numérique de la Bibliothèque nationale et universitaire de Strasbourg), constitue une sorte d'équivalence aux portails institutionnels offerts par HAL. Au sein de ce dispositif, les données sont conservées à la BnF, et bénéficient d'une solution d'archivage pérenne, le Système Pérenne d'Archivage Réparti[66] (SPAR). Les institutions qui choisissent cette solution ont ensuite une bibliothèque numérique où les données sont rapatriées depuis les serveurs de Gallica, et bénéficient des mêmes fonctions de recherche, de visualisation et d'éditorialisation des collections. La logique de recentralisation des contenus, plutôt que de l'échange des métadonnées qui les concerne, est donc également à l'œuvre dans le domaine du patrimoine. La transition est importante : nous passons d'un système centré autour des métadonnées à un système centré autour de la ressource. Dans ce contexte, on peut à nouveau questionner l'utilisation du protocole OAI-PMH : pourquoi ne pas transmettre directement les données, accompagnées de métadonnées plus riches, à ces quelques partenaires intéressés ?

[64] Les exemples retenus, et de nombreux autres, sont présentés dans : BIBLIOTHÈQUE NATIONALE DE FRANCE, Direction des Services et des Réseaux, Département de la Coopération. *Les partenaires de Gallica en 2017* [en ligne]. 2017. [Consulté le 14 février 2020]. Disponible à l'adresse : https://www.bnf.fr/sites/default/files/2018-11/partenaires_gallica.pdf.

[65] Gallica marque blanche. Dans : *BnF - Site institutionnel* [en ligne]. [s. d.]. [Consulté le 14 février 2020]. Disponible à l'adresse : https://www.bnf.fr/fr/gallica-marque-blanche.

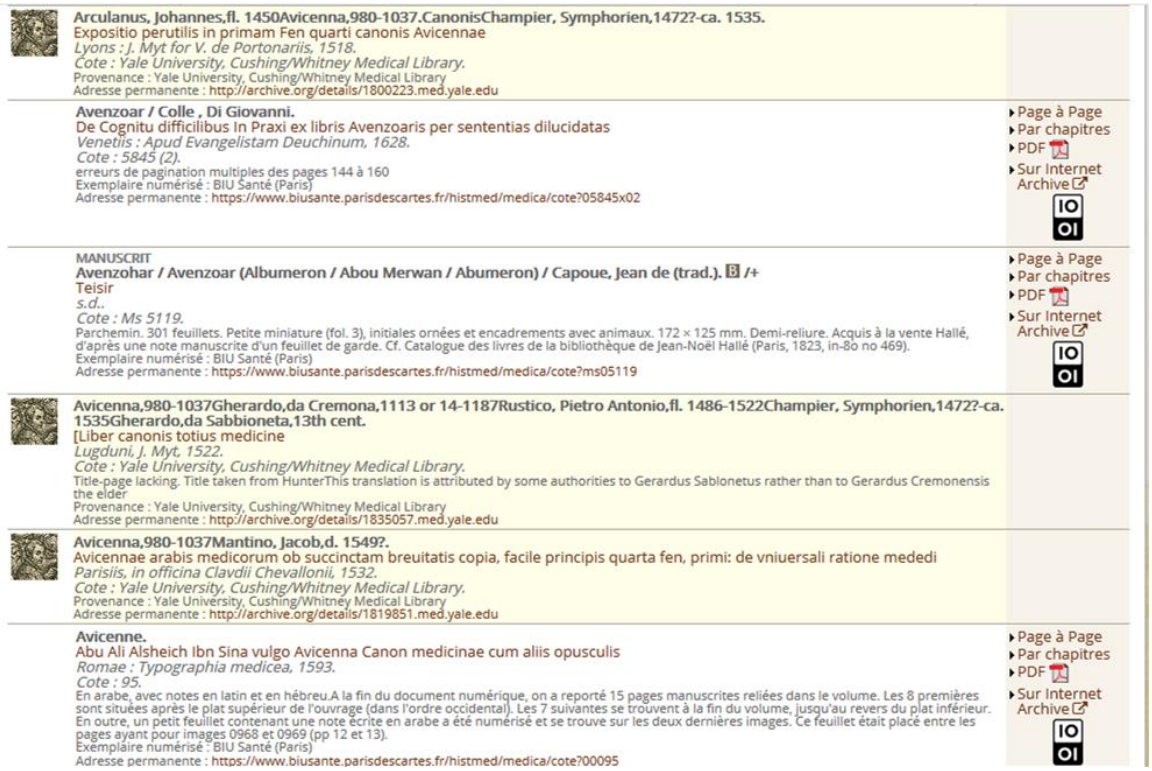
[66] SPAR (Système de Préservation et d'Archivage Réparti). Dans : *BnF - Site institutionnel* [en ligne]. [s. d.]. [Consulté le 14 février 2020]. Disponible à l'adresse : https://www.bnf.fr/fr/spar-systeme-de-preservation-et-darchivage-reparti.

Dublin Core : les limites de l'élasticité

L'élasticité du Dublin Core a définitivement démontré ses limites, puisqu'un mapping est nécessaire pour chaque entrepôt rencontré. Dans ce contexte, on peut même questionner la notion même d'interopérabilité du protocole : si chaque utilisation demande un mapping, pourquoi ne pas transporter directement des données dans un format plus riche, et concevoir un mapping depuis ce format ? De manière générale, il faut noter que le Dublin Core n'est plus, voire n'a jamais été, un langage commun à toutes les institutions qui l'utilisent. À l'heure actuelle, il existe plusieurs Dublin Core, et les institutions comme Gallica qui moissonnent des entrepôts dans le cadre du protocole ont une interprétation de son utilisation qui fait partie des éléments du partenariat.

De manière générale, les collaborations qui utilisent aujourd'hui le protocole OAI-PMH se font dans le cadre de partenariats de gré à gré. Techniquement, rien n'empêche le moissonnage d'un entrepôt sans ce partenariat, les données qui sont contenues à l'intérieur étant librement accessibles par quiconque suit les spécifications du protocole. Mais cette utilisation « sauvage » ne semble pas répandue, au moins dans les institutions contactées dans le cadre de ce travail, ni dans les exemples glanés dans la littérature professionnelle. Aujourd'hui, on pourrait imaginer d'autres manières de transmettre des données entre institutions. Mais abandonner totalement OAI-PMH pour un autre protocole poserait également problème : si d'autres solutions existent, sont-elles aussi simples à implémenter ? Ont-elles une plasticité équivalente ? L'utilisation du protocole par la BIUS ne correspond pas au cas d'usage imaginé par les personnes ayant conçu le protocole, mais bien à l'usage qu'en a fait la communauté patrimoniale. En effet, la contractualisation, la participation à des initiatives internationales comme la Medical Heritage Library ou Europeana, correspondent à l'essaimage d'OAI-PMH dans le milieu des années 2000. C'est aujourd'hui un outil connu, sur lequel les professionnel-le-s ont un véritable retour réflexif. Si l'on en connaît les limites, il faut également reconnaître ses avantages : il a poussé un nombre très important d'institutions à exposer leurs données sur le web. Si le Dublin Core porte son lot de problématiques, liées à sa plasticité, il est toutefois assez simple à prendre en main, et, s'il est moins riche, il est aussi moins compliqué que d'autres formats « métier » comme les différentes formes de MARC. Mais, comme on l'a vu, dans l'idée des concepteurs d'OAI-PMH, les éléments du Dublin Core n'étaient qu'un début. L'absence d'ajouts d'autres métadonnées condamne en quelque sorte le protocole à rester tel que nous le voyons implémenté actuellement. L'exemple de Medica est assez riche : dans une liste de résultats,

comme celle présentée ci-dessous, on voit apparaître les documents de la BIUS avec plusieurs options de consultations (page à page, par chapitres), une indication claire des droits associés (licence Etalab).



Capture d'une page de résultats de Medica[67]

Les documents moissonnés avec OAI-PMH, nous pouvons les reconnaître grâce à une vignette qui apparaît à gauche du titre, mais aussi à travers l'absence de ces options de consultation riches. Les métadonnées qui apparaissent, de manière assez désordonnée en raison de l'utilisation du Dublin Core (le champ auteur comporte de manière indiscriminée des noms, dates, qui plus est selon des formats et des langues différentes) ne comportent pas les droits, mais seulement un champ auteur, des champs de titre/sous-titres, une cote (encore une fois, non standardisée : elle apparaît une fois de manière numérique, une autre fois avec une chaîne de texte indiquant différentes institutions...), une provenance et une adresse permanente, c'est-à-dire un URL auquel la ressource peut être consultée. Par rapport à la solution technique que représente l'OAI-PMH pour le transport de données, qu'est-il possible de faire à travers l'utilisation d'un autre schéma d'organisation de données ?

[67] Correspondant à la recherche suivante : *« Médecine arabe »— Medica — BIU Santé, Paris* [en ligne]. [s. d.]. [Consulté le 14 février 2020]. Disponible à l'adresse : https://www.biusante.parisdescartes.fr/histoire/medica/resultats/?intro=arabe&statut=charge.

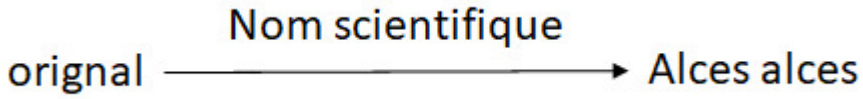
LE WEB SEMANTIQUE : DES PROMESSES A LA REALITE

L'infrastructure : lier les données entre elles

Au moment où OAI-PMH apparaît, en 2001, Tim Berners-Lee, James Hendler et Ora Lassila[68] publient un article dans le *Scientific American* qui décrit deux personnes qui, à l'aide de navigateurs web tenus à la main (*handheld browsers*), prennent un rendez-vous médical pour leur mère malade. La particularité de cette prise de rendez-vous est que les appareils des deux personnes communiquent, entre eux et avec d'autres, afin de transformer la requête orale en langage compris par la machine. Les mots utilisés par les humains sont compris par la machine, par le biais des technologies du web dit sémantique. En somme, une chaîne de caractères, comme le mot « orignal », n'est plus interprétée par l'ordinateur comme une suite de caractères binaires (en l'occurrence, « 01101111 01110010 01101001 01100111 01101110 01100001 01101100 »), mais par un Identifiant de Ressource Uniforme (*Uniform Resource Identifier*, ou URI) qui permet à la machine de comprendre qu'un orignal est un mammifère ruminant de la famille des cervidés. En comprenant la requête, la machine renvoie alors des résultats qui répondent à la demande originale. Si ce modèle de description de données est généralisé, la machine peut aller rechercher l'information à plusieurs endroits et ainsi répondre à des requêtes complexes. Par exemple, afin de savoir combien d'orignaux vivent en Amérique du Nord, la machine pourrait aller chercher dans les bases de données des agences de protection de l'environnement des pays de cette zone géographique pour rechercher l'information, tout en ayant compris qu'un orignal, en anglais, était connu sous le terme *moose*. En somme, plutôt que de trouver une chaîne de texte qui ressemble à la recherche, la machine trouve la signification de cette chaîne. La promesse du web sémantique est donc de capitaliser sur la décentralisation qui est à l'origine du World Wide Web, et de transformer la manière dont l'information est indiquée, et donc la manière dont la machine la comprend.

Cela se fait à travers la création de liens, dans le modèle *Resource Description Framework* (RDF). Par exemple, on indiquera :

[68]BERNERS-LEE, TIM, HENDLER, JAMES et LASSILA, ORA. THE SEMANTIC WEB. *Scientific American*. 2001, Vol. 284, nº 5, p. 34-43. JSTOR.



Ce lien est un triplet, et chaque élément en son sein a, dans ce modèle, un URI, qui permettra à la machine de le différencier. On peut donc lier, de cette manière, une infinité d'informations, et même de documents : toujours dans l'exemple de notre orignal, il est possible de lier une photographie, par exemple, en précisant par le lien que la photographie représente l'orignal.

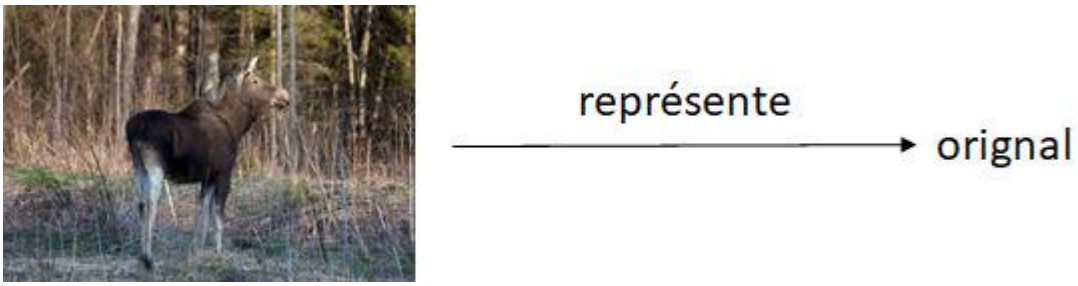


Photo : Villa16, CC BY SA 3.0 (https://commons.wikimedia.org/wiki/File:Los.jpg)

Pour en revenir à des problématiques bibliothéconomiques, dans un tel modèle, on pourrait ainsi lier une thèse à un exemplaire physique, à un substitut numérique, la personne qui l'a écrite et la personne qui l'a dirigée en indiquant précisément les relations. C'est un modèle conçu dans l'écosystème du Web, et qui va donc beaucoup plus loin que les ressources documentaires : en somme, utilisé comme dans l'exemple de Berners-Lee, Lassila et Hendler, il permet de décrire des documents avec une précision supérieure à MARC, et avec une élasticité supérieure à celle de Dublin Core.

De la création du graphe à son exploration

La description de contenus par la création de liens n'est que la première étape nécessaire à la création de services en données liées. Elle correspond à notre logique bibliothéconomique traditionnelle, puisque créer ces liens, c'est décrire le plus précisément possible une chose (document, contenu, espèce vivante, théorie...). Mais à l'inverse de nos technologies de description actuelles, les technologies en données liées permettent

d'exploiter réellement la richesse de nos descriptions : la recommandation Simple Knowledge Organization System[69] (SKOS), publiée en 2009 par le World Wide Web Consortium (W3C), permet par exemple de représenter des langages documentaires, thésaurus et classifications, afin d'en faire des vocabulaires structurés utilisable dans le web sémantique ; l'ontologie Friend Of A Friend[70] (FOAF), permet de décrire finement les personnes (nom, prénom, surnom, date de naissance…), et les liens entre personnes et entités (membre de, connaît telle personne, basé près de). Ces deux exemples, combinés, permettent d'utiliser des données de nos bibliothèques, comme les référentiels, et de les exprimer en données liées. Ces deux ontologies ne sont qu'une petite partie des centaines différentes dans le monde du web sémantique, générales ou extrêmement spécialisées (par exemple en biologie, astrophysique…), qui permettent d'avoir une description fine et enrichie d'un élément. Ainsi, on peut décrire aussi bien les relations sociales d'une personne que sa classificiation biologique, sa localisation géographique, ses possessions, son éducation, son activité professionnelle, et bien plus encore. De plus, les descriptions viennent à se croiser : par exemple, deux humains auront la même classification biologique, mais des localisations géographiques différentes, alors qu'un animal pourra avoir la même localisation géographique qu'un humain mais pas la même classification biologique, et ainsi de suite.

La manière dont cela fonctionne demande d'étendre l'exemple que nous venons de présenter. À supposer que nous continuions à décrire un orignal, nous aboutirions à un graphe qui ressemblerait à ceci.

[69] Pour voir les différentes classes, consulter l'espace de nommage : *SKOS Simple Knowledge Organization System Namespace Document - HTML Variant, 18 August 2009 Recommendation Edition* [en ligne]. [s. d.]. [Consulté le 22 février 2020]. Disponible à l'adresse : https://www.w3.org/2009/08/skos-reference/skos.html#mappingRelation.

[70] Voir les spécifications sur : *FOAF Vocabulary Specification* [en ligne]. [s. d.]. [Consulté le 22 février 2020]. Disponible à l'adresse : http://xmlns.com/foaf/spec/#term_age.

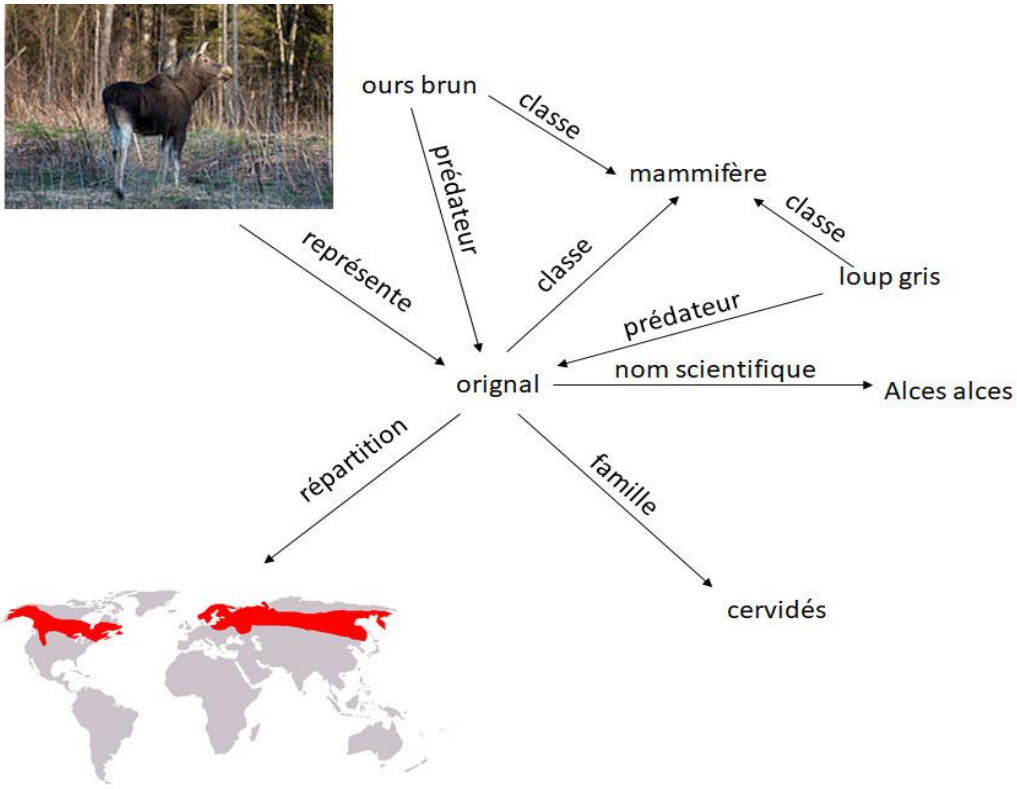


Figure 1 : graphe étendant la description de l'orignal[71]

En continuant de tisser des liens de ce genre, nous aboutirions à un graphe encore plus complexe, et, de manière rapide, illisible ou inutilisable par l'humain. C'est donc là qu'intervient le protocole de requêtage, SPARQL (*SPARQL Protocol And RDF Query Language*). Grâce à SPARQL, on explore le graphe de données que l'on a créé. Dans l'exemple ci-dessus, par exemple, on pourrait demander, via une requête, une liste de tous les animaux dont la « classe » est « mammifère », et la machine retournerait donc l'ours brun, le loup gris et l'orignal. Si l'on faisait une requête pour avoir une liste de tous les « prédateurs » de l'orignal, la machine retournerait le loup gris et l'ours brun. Mais les données de répartition, par exemple, sont géographiques. Si elles sont renseignées avec des coordonnées GPS, on peut avoir des résultats sous forme de cartes ; si les données sont temporelles, on peut alors avoir des résultats sous forme de ligne temporelle...de la même manière, si les images sont convenablement liées, les recherches d'image se feront avec moins de bruit dans les résultats. L'intérêt principal du requêtage est que le protocole est extrêmement souple, et peut être adapté aux usages que veulent en faire les personnes. On

[71] Photographie de l'orignal : Villa16, CC-BY-SA 3.0 (https://commons.wikimedia.org/wiki/File:Los.jpg) ; carte de la répartition : GBRUICKER Jürgen, CC-BY-SA 3.0 (https://commons.wikimedia.org/wiki/File:Moose_distribution.png)

peut l'utiliser aussi bien pour exploiter des données géographiques que temporelles, picturales que textuelles. Contrairement à OAI-PMH, SPARQL est un protocole synchrone : l'interrogation se fait en temps réel, c'est à dire que si des données sont rajoutées à mon graphe, la même requête, faite quelques secondes plus tard, renverra plus de résultats. C'est donc un protocole dynamique, mais plus lourd que l'asynchronie d'OAI-PMH que nous avons évoquée. En effet, la requête SPARQL peut explorer plusieurs bases de données (à travers la recherche fédérée) : elle agit en repérant les contenus recherchés dans plusieurs bases, décentralisées. Cette logique correspond à la construction du web, à sa structure par essence décentralisée. Les requêtes recentralisent cette information à la demande d'un usager.

Des promesses dans le monde des GLAM

Pour la communauté des GLAM, les technologies du web sémantique offrent des promesses intéressantes. Elles répondent, comme OAI-PMH, au besoin de rendre l'information accessible sur le web, mais d'une manière infiniment plus précise. Elles respectent une architecture décentralisée, et permettent d'imaginer des utilisations nouvelles de leurs contenus : c'est une manière de rendre des documents, des contenus accessible à des communautés d'usager-e-s, qui peuvent les utiliser dans des modalités nouvelles et inattendues. OAI-PMH, même s'il permet de transporter plus que de « simples » métadonnées, est majoritairement utilisé pour ne transmettre que ces dernières ; à travers le web de données, on donne accès directement aux contenus. Le processus de requête expliqué plus haut permet de sortir l'information des différents silos propres aux communautés professionnelles, et de la rendre accessible par l'usager. Ce que l'on entend par le terme « silo » est le stockage de l'information dans une base locale, isolée. En bibliothèque, le catalogue est un silo : c'est la base où les agent-e-s stockent la description des contenus qui sont présents dans l'établissement. Cette base est accessible aux lecteurs et lectrices, *via* un moteur de recherche qui ne cherche que dans ce catalogue. Mais le contenu de cette base n'est pas accessible pour les moteurs de recherches comme Google, Yahoo ou Qwant : ce contenu est dans le web profond (*deep web*), c'est-à-dire non référencé dans le moteur de recherche. Le catalogue de bibliothèque, par exemple, est un silo, celui des archives en est un autre : le défi, pour nos professions, consiste à exposer nos contenus sur le web, en-dehors de ces silos où ils ne peuvent être découverts.

Là où OAI-PMH nécessite d'abord une exposition, puis un moissonnage avant de donner accès à des données, le requêtage SPARQL permet d'avoir directement à des données correctement exposées. Du point de vue des professionnel-le-s de l'information, cela permet de donner accès à un document, mais aussi à tout le contexte qui l'entoure.

Dans le cas d'un roman, par exemple : si un lecteur cherche, dans un catalogue en ligne, un roman de Boris Vian, il pourra également avoir accès à ceux de Vernon Sullivan (pseudonyme de l'auteur), à ses enregistrements de jazz, aux adaptations cinématographiques de ses travaux, mais aussi au collège de Pataphysique (dont l'auteur était membre), voire les brevets qu'il a déposés, car tous ces contenus seront liés à leur créateur de la même manière que les exemples que nous avions pris. En données liées, les contenus ne sont pas, d'ailleurs, nécessairement propres au catalogue de l'institution. Une requête SPARQL cherche, dans les critères qui lui ont été donnés, tous les éléments qui correspondent dans les bases de données qu'on indique. Ainsi, le web sémantique permet d'exposer des données, des contenus d'une manière beaucoup plus riche qu'à l'heure actuelle, et permet d'encourager des collaborations transdisciplinaires et le croisement entre professions de la culture ou de la recherche. De ces promesses est né un véritable mouvement d'évangélisation des technologies du web sémantique, qui existe depuis plusieurs années, en France comme dans le reste du monde. On peut le voir par le nombre d'ouvrages disponibles sur la question, qui apparaissent surtout à partir de la fin des années 2000, principalement en langue anglaise[72], mais aussi en français, avec la publication d'Emmanuelle Bermès, Antoine Isaac et Gautier Poupeau[73]. À ces publications, il convient d'ajouter l'activité importante de la blogosphère, qui représente également un espace où des professionnel-le-s ont communiqué sur les possibilités qu'offrait le web sémantique dans les métiers des professionnel-le-s de l'information, sur des retours d'expérience et des questionnements. Il est important de rappeler ce contexte, qui explique l'enthousiasme qu'ont pu soulever ces technologies. Dans un sens, il faut lire cette effervescence en parallèle avec la communauté des professionnel-le-s qui ont contribué à créer et diffuser le protocole OAI-PMH. Pendant un temps, celui-ci a suscité de l'enthousiasme et de l'engagement dans les communautés professionnelles, ce qui a aussi été le cas pour les technologies du web sémantique. Pour rendre compte des réflexions et des

[72]On peut par exemple citer : WELLER, Katrin. *Knowledge representation in the social semantic Web*. Berlin New York : De Gruyter Saur, 2010. ISBN 978-3-598-25180-1. TK5105.88815 .W45 2010 ; HYVÖNEN, Eero. *Publishing and using cultural heritage linked data on the Semantic Web*. San Rafael, Etats-Unis d'Amérique : Morgan & Claypool, cop 2012. ISBN 978-1-60845-998-8.

[73]BERMES, Emmanuelle, ISAAC, Antoine et POUPEAU, Gautier. *Le web sémantique en bibliothèque*. Paris, France : Éd. du Cercle de la librairie, 2013. ISBN 978-2-7654-1417-9.

réalisations qui découlent de cet enthousiasme, il convient de prendre un exemple : celui d'OAI-ORE nous permet de faire le lien entre OAI-PMH et le concept de web sémantique.

OAI-ORE : une succession non advenue ?

Nous avons déjà analysé la réaction des personnes responsables de la création et du développement d'OAI-PMH, et leur déception face à l'absence d'enrichissements des données utilisées. La réponse a été relativement rapide : pendant que le protocole essaime dans les milieux de la culture et de la recherche dans la seconde partie des années 2000, Carl Lagoze et Herbert Van de Sompel, avec quelques autres contributeurs, travaillent à la création d'un nouveau protocole. Conscients des limites du précédent, le nouveau protocole est conçu comme une manière d'appréhender les objets composés (*compound objects*), c'est-à-dire des contenus composés de plusieurs ressources, et d'exprimer les liens qui leur sont propres, afin d'éviter de reproduire un échange « plat » d'information, comme c'est le cas à travers OAI-PMH.

Décrire des objets complexes

Il convient de comprendre ce que l'équipe de développement entend par des objets composés :

> " Quelques exemples de [ces objets composés] pourraient être un livre numérisé, qui est un assemblage de chapitres, où chaque chapitre est un assemblage de pages scannées ; un album de musique qui est un assemblage de plusieurs pistes audio ; une image d'objet qui est un assemblage d'une épreuve de haute qualité, un dérivatif de qualité moyenne et une vignette de basse qualité ; une publication scientifique qui est un assemblage de texte et d'autres matériaux tels que des jeux de données, des logiciels ou des enregistrements vidéo d'une expérience ; un document sur le web, en plusieurs pages, avec une table de contenus HTML qui indique plusieurs pages HTML liées entre elles. "[74]

Il est intéressant de noter que le choix de ces documents complexes, et encore assez neufs, semble répondre aux problématiques auxquelles essayait de répondre le groupe de travail ayant élaboré le Dublin Core : trouver une manière de décrire des contenus qui deviendront, avec les objets nativement numériques, plus protéiformes et complexes. Quelques années après la création de Dublin Core, la Bibliothèque du Congrès crée un autre standard afin de décrire un document avec un degré de finesse important : le Metadata Encoding and

[74]"Some examples of these are a digitized book that is an aggregation of chapters, where each chapter is an aggregation of scanned pages; a music album that is the aggregation of several audio tracks; an image object that is the aggregation of a high quality master, a medium quality derivative and a low quality thumbnail; a scholarly publication that is aggregation of text and supporting materials such as datasets, software tools, and video recordings of an experiment; and a multi-page web document with an HTML table of contents that points to multiple interlinked HTML individual pages", traduction personnelle de ALLINSON, Julie. *Thoughts on Compound Documents* [en ligne]. mai 2005. [Consulté le 2 février 2020]. Disponible à l'adresse : http://www.openarchives.org/ore/documents/CompoundObjects-200705.html.

Transmission Standard[75] (METS). Ce format comporte sept parties distinctes : un en-tête, un champ de métadonnées descriptives, un champ de métadonnées administratives, une liste des fichiers, une carte de la structure, un champ de liens structurels, et une partie expliquant le comportement. De cette manière, on peut exprimer de manière plus précise la structuration d'un contenu, ce qui peut être utile notamment dans le champ des archives, où ce standard est très utilisé. Mais ce standard, tout bien installé dans les communautés professionnelles, et bien que compatible avec le protocole OAI-PMH, ne remplace pas le Dublin Core, probablement en raison d'une certaine complexité, qui va de pair avec la précision et la granularité qu'il offre.

Dans le contexte, la problématique demeure en 2007, particulièrement avec un premier retour d'expérience d'OAI-PMH : on peut techniquement décrire des objets complexes avec Dublin Core, mais comment faire pour ne pas les « séparer » en les exposant avec le protocole ? Comment conserver les liens entre différentes ressources ? À travers OAI-PMH, cela est pratiquement impossible, comme le rappelle Gautier Poupeau en 2009 :

> « Même s'il est effectivement possible de relier les ressources décrites au sein d'un entrepôt OAI-PMH par l'intermédiaire des métadonnées, ces dernières sont encapsulées au sein d'un record. Or, chaque record est **indépendant** et forme une représentation documentaire de la ressource décrite sous la forme d'une notice. Ainsi, lorsque chaque record sont indexés par un service provider, cette notion tend à mettre au même niveau tous les types de ressources. Si vous faites une recherche sur OAISTER, tout est au même niveau, article, ouvrage, thèse, chapitre d'ouvrage, numéro de revue voire différents niveaux de description archivistique, il n'est ainsi pas possible au sein du protocole (je parle bien du protocole et pas des métadonnées qui y sont exposées) d'exprimer la granularité et les relations. »[76]

La problématique est donc celle de la persistance d'une manière ancienne, et très bibliothéconomique, de concevoir la description de l'information, c'est à dire : un document, une notice. Exposer des documents sur le web pour qu'ils soient exploités au maximum de leurs possibilités, c'est prévoir une description plus fine, et surtout une description adaptée à des contenus qui sont apparus après cette manière de classer l'information, comme certains exemples donnés plus hauts. Il faut également prendre en compte que, depuis la fin des années 1990, une partie de la communauté des professionnel-le-s de l'information cherche à se séparer de l'approche traditionnelle de description de l'information : en 1998 apparaissent les Functional Requirements for Bibliographic Records[77] (FRBR), développés par

[75] *Metadata Encoding and Transmission Standard (METS) Official Web Site | Library of Congress* [en ligne]. [s. d.]. [Consulté le 22 février 2020]. Disponible à l'adresse : https://www.loc.gov/standards/mets/.

[76] POUPEAU, Gautier. *Les carcans de la pensée hiérarchique et documentaire (2) | Les petites cases* [en ligne]. [s. d.]. [Consulté le 31 janvier 2020]. Disponible à l'adresse : https://www.lespetitescases.net/carcans-de-la-pensee-hierarchique-et-documentaire-2.

[77] *IFLA -- Functional Requirements for Bibliographic Records* [en ligne]. [s. d.]. [Consulté le 22 février 2020]. Disponible à l'adresse : https://www.ifla.org/publications/functional-requirements-for-bibliographic-records.

l'International Federation of Library Associations (IFLA), qui introduit quatre niveaux de description des contenus : œuvre, expression, manifestation, item. Avec ce modèle, on change de la vision à l'exemplaire, pour rentrer dans un processus de description plus complexe, qui correspond à l'arrivée des formats numériques, hybrides, à différentes manifestations ou expressions d'une même œuvre.

Dans le contexte de documents complexes, particulièrement dans le milieu de le recherche, il convient donc de trouver une solution pour exprimer les liens entre différents documents :

- Pour des documents qui perdent leur richesse ou leur pertinence en étant « mis à plat », comme des revues (qui sont composées d'un ensemble de numéros) ou des fonds d'archive (avec plusieurs niveaux de description)

- Pour des documents nativement numériques (pages web) ou des contenus faisant usage de technologies du web afin d'exprimer, en leur sein, plusieurs autres liens (hypertextes, par exemple)

Pour répondre à ces problèmes, Lagoze et Van de Sompel conçoivent l'Open Archives Initiative – Object Reuse and Exchange (OAI-ORE), un système qui se repose sur les technologies du web sémantique afin d'exposer l'information tout en respectant ses liens internes. Tout comme dans la logique du web sémantique, il s'agit d'indiquer les liens entre différents objets, et de qualifier ces liens pour les rendre intelligibles.

De la métadonnée à la ressource

Contrairement à OAI-PMH, il n'y a pas d'adoption d'un format particulier en termes de métadonnées, sachant que les deux projets sont fondamentalement différents : le premier était centré sur les métadonnées qu'il permettait d'échanger, alors que le second se concentre sur les ressources. Ainsi, OAI-ORE n'est pas conçu afin de remplacer OAI-PMH : ce n'est pas une version ultérieure, nouvelle du protocole, mais une approche différente afin de répondre à un problème semblable, la visibilité des ressources et leur appréhension sur le web. Au lieu de recentraliser les données dans un entrepôt, OAI-ORE respecte la décentralisation des contenus, pour coller à l'architecture du web, et toujours dans la logique du web sémantique. Ce nouveau projet, qui se définit comme une spécification plutôt qu'un protocole, cherche à concevoir un modèle qui permette de présenter des objets complexes, tout en respectant leur particularité. C'est donc une démarche fondamentalement différente

d'OAI-PMH, qui nécessitait, au minimum, une transformation des données selon les éléments du Dublin Core, un élément qui, on l'a vu, pose des problèmes pour la qualité des données transmises. Ici, OAI-ORE se veut comme une couche d'interopérabilité (*interoperability layer*) mise sur des objets complexes, pour les rendre appréhensibles par la machine, et faciliter les opérations de recherche et de découverte des contenus. Dans cet environnement, il n'est pas nécessaire qu'OAI-PMH disparaisse ; il peut rester en usage pour des ressources qui ne posent pas ce problème de multiplicité d'état, et être complémenté par OAI-ORE. Mais dans la vision de l'avènement du web sémantique, qui devait amener une véritable révolution des usages, on peut supposer que, du point de vue des concepteurs et conceptrices d'OAI-ORE, l'utilisation de cette spécification allait contribuer à rendre obsolète le protocole : en parvenant à exprimer les liens entre les documents, en respectant une structure décentralisée, quelle serait l'utilité d'une opération de recentralisation de métadonnées pour visibiliser des ressources ? De la même manière que, par l'usage, OAI-PMH avait créé des usages, OAI-ORE avait un potentiel de transformation par l'usage, d'autant plus que le précédent protocole, avait bien essaimé. Le développement d'ORE était donc suivi par la communauté, et par des acteurs majeurs du numérique, notamment Microsoft et Google, qui faisaient partie du groupe de liaison dans le processus de développement. Microsoft, ainsi que la Mellon Foundation, a également financé une partie du développement de la spécification[78].

Le but d'OAI-ORE est donc de faire sens d'un contenu hétérogène ; le mécanisme servant à remplir cet objectif réunit quatre éléments, une représentation (*representation*), une carte des ressources (*resource map*), un assemblage (*aggregation*) et enfin les ressources elles-mêmes (*aggregrated* resources). Chacun de ces éléments est lié au suivant. La représentation renvoie à la carte des ressources, qui décrit l'assemblage, et ce dernier est composé des différentes ressources. Ainsi, la représentation restitue l'intégralité d'un assemblage, mais permet également d'en conserver la structure, accessible par les procédés de la négociation de contenus[79]. La négociation de contenus est un mécanisme de l'Hypertext Transfer Protocol (HTTP), qui permet de fournir différentes représentations d'une ressource à la même URI : une page web dans plusieurs langues, une image en plusieurs formats ; selon les capacités et

[78]Voir la composition du groupe de liaison et la déclaration de remerciement aux institutions finançantes dans ALLINSON, Julie. *Thoughts on Compound Documents* [en ligne]. mai 2005. [Consulté le 2 février 2020]. Disponible à l'adresse : http://www.openarchives.org/ore/documents/CompoundObjects-200705.html.

[79]Principe expliqué dans SAUERMANN, Leo, GMBH, Dfki et CYGANIAK, Richard. *Cool URIs for the semantic web*. Juillet 2011.

les réglages du navigateur, du serveur ou les choix de l'utilisateur, la ressource sera présentée sous la forme d'une représentation plutôt que d'une autre.

La conception de ce modèle permet donc d'exprimer pleinement la richesse, la structure et la complexité d'un contenu. Chaque élément est doté de métadonnées propres, qui peuvent être exprimées dans des vocabulaires riches du web sémantique : l'exemple donné lors de la première version de la spécification utilise, en dehors du Dublin Core, l'ontologie Friend of a Friend (FOAF) et le format de syndication Atom, qui permettent un niveau de détail et de variété de métadonnées important. Cette flexibilité est cruciale pour décrire des liens et des documents complexes : si les six ressources qui constituent un assemblage ont six auteurs, il convient de pouvoir faire les attributions correctes, et d'attribuer la bonne ressource à son auteur. Il n'y a donc pas, dans OAI-ORE, un plus petit commun dénominateur, mais un cadre dans lequel peut être développé un ensemble de liens.

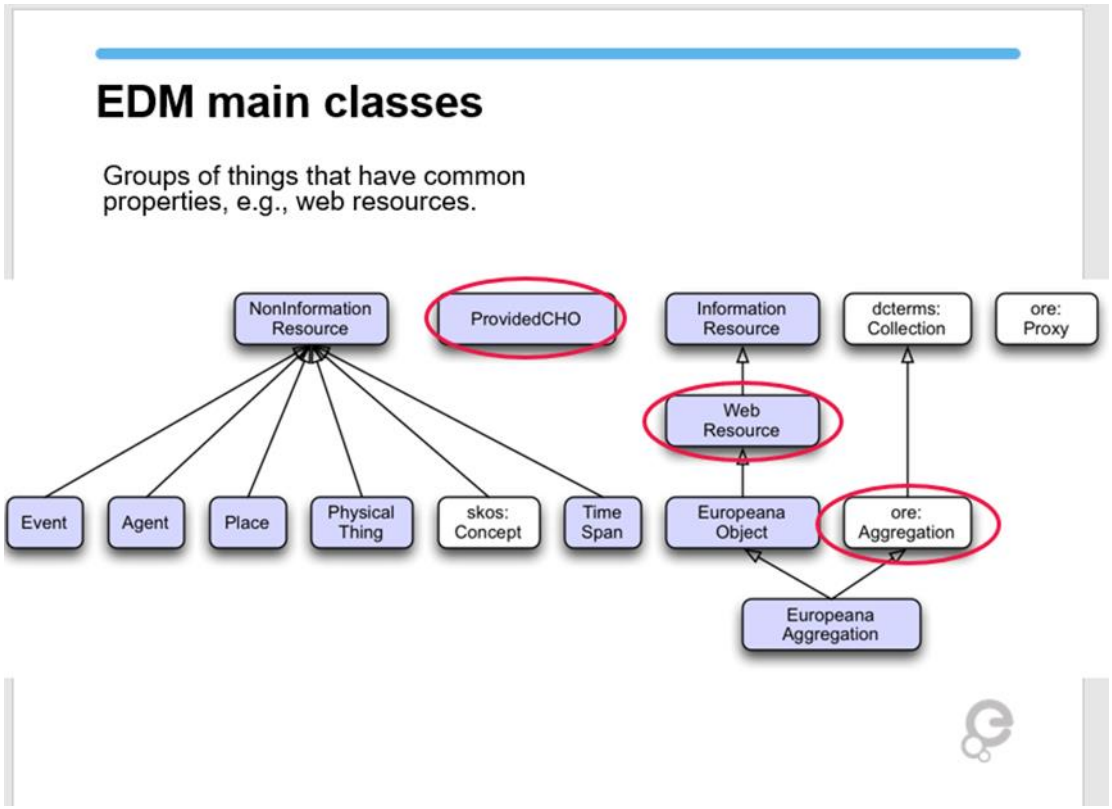
Un succès limité ?

Quelle a été la dissémination et les usages qui ont accompagné la définition d'OAI-ORE ? Il est aujourd'hui complexe de statuer sur cette dissémination, car peu d'implémentations sont documentées. À partir de 2008, et donc de la sortie de la première version de la spécification, on trouve sans peine des communications sur l'intérêt de la spécification, et quelques applications lors de projets : en 2008, un projet australien qui utilise ORE pour éditer des objets composés, dans le cadre d'un projet de littérature comparée[80] ; trois ans plus tard, la Bibliothèque du Congrès utilise la spécification dans le cadre d'un projet de portail de presse ancienne[81].

De manière générale, la diffusion de la spécification s'est faite de manière relativement confidentielle : on trouve ORE dans le modèle de données d'Europeana, mais de manière discrète : c'est une ontologie parmi d'autres. Il n'y a donc pas l'effet d'annonce qu'on a pu voir lors de l'utilisation d'OAI-PMH.

[80]GERBER, Anna et HUNTER, Jane. LORE: A compound object authoring and publishing tool for the Australian literature studies community. Dans : *International Conference on Asian Digital Libraries*. Springer, 2008, p. 246–255.

[81]THOMAS, Deborah. *Gathering History: Collaboration and the National Digital Newspaper Program*. 2011.



Extrait de la présentation de l'European Data Model (EDM) (CC-BY-SA)

L'Europeana Data Model remplace en 2010 le premier modèle de données d'Europeana, qui était appelé Europeana Semantic Elements (ESE), mais était un système plat, qui ne permettait pas de rediriger vers d'autres contenus, et de décrire plus finement les contenus. EDM permet donc à Europeana d'être « une grande agrégation de représentations digitales d'artefacts culturels, accompagnée d'une contextualisation riche, et intégrée dans une architecture en données liées et ouvertes »[82]. Dans le *primer,* qui est un document explicatif du modèle, on trouve un remerciement direct à Herbert Van de Sompel, pour des contributions, ce qui explique également l'utilisation d'ORE, qui a pu être présenté par son créateur. De la même manière qu'OAI-PMH a circulé de la BnF à Europeana, il semble qu'ORE ait fait le processus inverse, puisqu'on le retrouve dans le modèle de données de la BnF[83], afin de gérer le niveau « Œuvre », dans le modèle IFLA-LRM[84] : dans cette logique, il est en effet intéressant de réfléchir en termes de contenus agrégés et d'agrégations, et

[82]" a big aggregation of digital representations of culture artefacts together with rich contextualization data and embedded in a linked Open Data architecture", traduction personnelle de : DOERR, Martin, GRADMANN, Stefan, HENNICKE, Steffen, et al. The Europeana Data Model (EDM). *World Library and Information Congress: 76th IFLA General Conference and Assembly*. 2010, p. 10-15.

[83]Présentation du modèle de données. Dans : *data.bnf.fr* [en ligne]. [s. d.]. [Consulté le 10 février 2020]. Disponible à l'adresse : https://data.bnf.fr/opendata#Ancre5.

[84] *IFLA -- IFLA Library Reference Model (LRM)* [en ligne]. [s. d.]. [Consulté le 20 février 2020]. Disponible à l'adresse : https://www.ifla.org/publications/node/11412.

l'utilisation d'ORE est alors cette couche sémantique qui était l'objectif de Lagoze et Van de Sompel. Mais, alors que ces deux grandes institutions utilisent la spécification, on ne peut pas dire qu'elles aient participé à sa diffusion, ou du moins, pas d'une manière que l'on puisse retracer clairement.

On peut sans doute trouver d'autres cas d'usages, mais force est de constater que le succès de la spécification est très en-deçà de ce qu'il était possible d'attendre, après l'exemple d'OAI-PMH. Il faut pourtant noter que sa conception a été suivie de près par les différents acteurs nationaux et internationaux qui étaient impliqués dans les problématiques du web du point de vue des spécialistes de l'information. On voit par exemple qu'OAI-ORE a été présenté en 2007 dans une journée professionnelle du Groupe de Travail sur les Archives Ouvertes (GTAO), à l'initiative du consortium Couperin[85], donc pendant les phases de préparation de la spécification, il est évoqué en 2009 dans une présentation sur Sciences Po Institutional Repository (SPIRE), l'archive ouverte de Sciences Po Paris[86], toujours dans le cadre d'une journée d'étude de Couperin sur les archives ouvertes. Mais ces présentations ne semblent pas déboucher sur des utilisations.

Alors qu'OAI-PMH continue de se répandre au sein des GLAM dans les années 2010 et après, OAI-ORE ne parvient pas à se développer de la même manière, et, douze ans après sa création, on peine toujours à trouver des implémentations documentées, dans le monde des bibliothèques ou au-delà. OAI-PMH a été l'origine, ou tout au moins la solution technique de nombreuses collaborations, de nombreuses réalisations, dont on a cité certaines (Europeana, OAIster...) ; ORE, à l'heure actuelle, n'est pas cité comme un élément important dans une initiative marquante. En 2013 en France, le Ministère de la Culture conseille l'utilisation d'OAI-PMH pour les professionnel-le-s de l'information[87] : il est donc, douze ans après sa création, encore au centre des attentions d'une partie de la profession, là où ORE n'entre pas en considération. Il est évident que la diffusion d'une technologie peut être lente, ce qui explique qu'en 2013, OAI-PMH soit encore un sujet de discussion, voire une découverte pour des professionnel-le-s ; mais même en 2020, soit dans un laps de temps

[85]CONSORTIUM COUPERIN. *Groupe de Travail sur les Archives Ouvertes* [en ligne]. mai 2007. [Consulté le 3 février 2020]. Disponible à l'adresse : http://gtao.wikidot.com/journee-d-etude-21-05-07.

[86]LUTZ, Jean-François. *Spire : l'archive ouverte de Sciences Po* [en ligne]. 04:54:19 UTC. [Consulté le 3 février 2020]. Disponible à l'adresse : https://fr.slideshare.net/jflutz/spire-larchive-ouverte-de-sciences-po.

[87]NAWROCKI, François. *Patrimoine écrit - Le protocole OAI-PMH* [en ligne]. 2 mai 2013. [Consulté le 28 juin 2019]. Disponible à l'adresse : https://web.archive.org/web/20130502181820/http://www.patrimoineecrit.culture.gouv.fr/Num/OAI-PMH.html.

comparable à notre exemple, OAI-ORE n'est guère un sujet de conversation, en-dehors d'un cercle relativement restreint.

Plusieurs points sont à prendre en compte dans cette absence de succès : tout d'abord, ORE est plus complexe dans sa conception que ne l'était OAI-PMH. Il suppose une maîtrise des technologies et vocabulaires du web sémantique, là où le précédent ne demandait qu'une conversion en Dublin Core, au minimum (et ce minimum est, on l'a vu, souvent la totalité des données exposées). ORE permettait à chaque communauté d'utiliser des ontologies et des standards appropriés, plutôt que de faire converger tous les contenus vers un format unique. Dans l'exemple donné plus haut, Dublin Core, FOAF et Atom étaient pertinents à la description du contenu et des rapports de création, mais pour d'autres types de document, il aurait été tout à fait possible d'utiliser des ontologies différentes. Cette liberté est cruciale dans le cadre d'une spécification orientée sur les ressources, mais elle ne crée pas un dénominateur commun, comme OAI-PMH en possède avec le Dublin Core. Enfin, et surtout, OAI-ORE est conçu pour le web sémantique : il en utilise la logique, la structure et les standards. Sans doute l'absence de succès est due, dans une certaine mesure, à l'absence de l'avènement du web sémantique : la recentralisation de métadonnées dans un entrepôt, le moissonnage, ces parties restaient dans la logique bibliothéconomique, elle correspondait à des pratiques métiers. Si OAI-PMH était bien un protocole du web, il ne bouleversait pas les logiques professionnelles, et ne changeait pas spécifiquement l'appréhension des documents et de leur description. Sa dissémination ne s'est pas non plus effectuée du jour au lendemain : les institutions spécialisées qui étaient à peu près toutes impliquées dans sa création en ont été les pionnières. De grandes institutions nationales ont ensuite porté ce protocole, et chaque partenariat a engendré une adoption plus large d'OAI-PMH, à partir de ces grands foyers que peuvent être les bibliothèques nationales, notamment (on a vu l'exemple de la BnF et de la Bibliothèque du Congrès). À l'inverse, OAI-ORE se retrouve sporadiquement, mais sans une adoption franche, sans dissémination impulsée par de grands établissements. Pour expliquer cela, on peut postuler que l'intérêt du dispositif n'a pas été ressenti par les différentes communautés ; on peut aussi penser que le changement induit par ORE était plus important, et que les communautés professionnelles n'y étaient pas spécifiquement préparées. La préparation est d'ailleurs complexe, sachant que les standards et concepts du web sémantique n'ont que peu essaimé au-delà d'une communauté particulière. L'adoption par les entreprises et les grands acteurs du numérique (y compris les moteurs de recherche) aurait facilité cet

accompagnement au changement, en rendant chacun-e familier-e avec une manière différente de concevoir le web.

De cette manière, le cas d'OAI-ORE n'est pas tant la représentation d'une succession manquée d'OAI-PMH : les deux visent des buts différents, ne sont pas deux objets similaires (l'un est un protocole d'échange de données, l'autre une manière de décrire et de signaler des objets complexes) ; le cas d'OAI-ORE illustre plutôt un problème dans la diffusion du web sémantique, sans laquelle la prise en compte de dispositifs tels que celui-ci sera systématiquement moins virale. Dans ce cadre, il est important de questionner une notion qui est dans le titre même de ce travail, à savoir « l'heure du web sémantique », promise depuis le début des années 2000.

Quelles réalisations en 2020 ?

Les promesses du web sémantique sont à remettre dans un contexte très particulier, qui est celui d'un stade de développement qui semblait promis. En effet, le fait de théoriser le web social comme le web 2.0 se faisait avec la venue postérieure d'un web 3.0, tour à tour nommé web de données ou web sémantique. Ce web « social »[88] est théorisé comme un espace de communication, de construction de communautés plutôt qu'une manière d'échanger des documents et des données entre individus. À ce concept correspondent également des technologies particulières : wikis, flux RSS, applications web, qui structurent le web tel que nous le pratiquons aujourd'hui. Ces technologies, appliquées à travers des plate-formes comme Wikipédia, Facebook ou Twitter, ont transformé le web en profondeur. Le développement de communautés virtuelles[89], et l'implication forte des internautes dans ces communautés, ont poussé les entreprises à capitaliser sur l'aspect social du web, se désintéressant sans doute de l'évolution ultérieure qu'est le web sémantique, dont les standards ne se sont pas imposés : si la majeure partie des internautes ont déjà utilisé un wiki, lu un blog, combien ont utilisé SPARQL, même de manière inconsciente ? Le développement de moteurs de recherches sémantiques, qui aurait pu être une nouvelle étape dans l'évolution de l'accès à l'information, est au point mort : sur les vingt-trois moteurs de recherche

[88]Tel que décrit dans RHEINGOLD, Howard. *The Virtual Community: Homesteading on the Electronic Frontier*. [S. l.] : MIT Press, 23 octobre 2000, p. 334. ISBN 978-0-262-26110-4. Google-Books-ID: fr8bdUDisqAC.

[89]Observée et analysée depuis le milieu des années 2000, voir par exemple GUILLOU, Benjamin. *Le développement des communautés virtuelles ou réseaux sociaux - CREG* [en ligne]. 9 décembre 2008. [Consulté le 6 février 2020]. Disponible à l'adresse : https://creg.ac-versailles.fr/le-developpement-des-communautes-virtuelles-ou-reseaux-sociaux.

sémantiques répertoriés[90] sur le wiki du World Wide Web Consortium (W3C), cinq seulement existent encore, et les exemples de requêtes, supposées démontrer la plus-value de la recherche à travers le web sémantique, ne fonctionnent que sur deux d'entre eux, à savoir LodView et LodLive[91]. Ce wiki démontre d'ailleurs autre chose, au-delà de l'aspect expérimental jamais dépassé de ces technologies : sa dernière mise à jour remonte à plus de 11 ans. En effet, le W3C, garant de l'interopérabilité du web, n'est plus occupé de la même manière par le développement du web sémantique. À travers ses groupes de travail, tout au long des années 2000, le W3C a mis en avant les technologies, les standards du web sémantique, mais la direction prise par le web a fait émerger d'autres problématiques. La centralisation autour de grandes plates-formes, notamment les réseaux sociaux, pose par exemple des problèmes de gestion des données personnelles (illustrés notamment par le scandale Cambridge Analytica[92]). Les problématiques du Big Data et de l'intelligence artificielle qui ont émergé dans les dernières années sont également au cœur des préoccupations, et donc des travaux, du W3C. Dans cet environnement, le web sémantique a pris une place bien moins importante : au mieux, il est au service des développements futurs. La vision du W3C est celle des professionnels : il regroupe organismes de recherche, industriels et éditeurs informatiques, des entreprises et des opérateurs de réseaux. Elle est donc tournée vers les évolutions à venir. Pour interroger les réalisations du web sémantique, il importe également de se tourner vers les utilisateurs et utilisatrices, et d'observer les usages actuels.

Aujourd'hui, toute utilisation du web est tournée vers la contribution. La vague sociale du web a également contribué à recentraliser les lieux de l'information, qui sont ceux de la contribution. Dans le monde de l'information, Wikipédia[93], par exemple, centralise une masse d'informations qu'on aurait préalablement trouvée disséminée dans de nombreux sites web spécialisés. Les réseaux sociaux sont également un lieu de la recentralisation, à comparer, par exemple, avec la blogosphère. Nous sommes donc, à l'heure actuelle, dans un web centralisé, tourné autour de grands pôles.

---

[90]W3C. *Category:Semantic Web Browser - Semantic Web Standards* [en ligne]. 2009. [Consulté le 6 février 2020]. Disponible à l'adresse : https://www.w3.org/2001/sw/wiki/Category:Semantic_Web_Browser.

[91] — *LodView, giving data a new shape* [en ligne]. [s. d.]. [Consulté le 22 février 2020]. Disponible à l'adresse : https://lodview.it/ ; *LodLive - browsing the Web of Data* [en ligne]. [s. d.]. [Consulté le 22 février 2020]. Disponible à l'adresse : http://en.lodlive.it/.

[92]Cambridge Analytica : 87 millions de comptes Facebook concernés. *Le Monde.fr* [en ligne]. 4 avril 2018. [Consulté le 6 février 2020]. Disponible à l'adresse : https://www.lemonde.fr/pixels/article/2018/04/04/cambridge-analytica-87-millions-de-comptes-facebook-concernes_5280752_4408996.html.

[93]*Wikipedia* [en ligne]. [s. d.]. [Consulté le 23 janvier 2020]. Disponible à l'adresse : https://www.wikipedia.org/.

Cela signifie qu'en 2020, la pertinence de l'architecture d'OAI-PMH peut être questionnée : la logique d'une multiplicité de points d'accès, avec des services construits selon les usages, n'est plus forcément adaptée. Même les fournisseurs de service qui tentent de moissonner tous les entrepôts OAI-PMH n'y parviennent pas réellement : en 2017, une recherche a montré que les méta-catalogues tels qu'OAIster avaient des points aveugles en terme de données exposées, et qu'un peu moins de la moitié (42,3%) des entrepôts OAI-PMH étaient spécifiques à un méta-catalogue[94]. Cela signifie que pour avoir accès à des informations contenues dans des entrepôts, les usager-e-s doivent avoir la volonté, et la possibilité (en termes de connaissances de recherche documentaire) de chercher potentiellement dans plusieurs bases de données. À un moment où des moteurs de recherche universitaires très puissants, comme Google Scholar, donnent de très bons résultats, l'idée d'utiliser consécutivement quatre sites, comme OAIster, le Directory of Open Access Journal et Open Archives est assez contre-intuitive.

La recentralisation de l'information va de pair avec cette performance des moteurs de recherche : puisque ces derniers sont plus performants, et plus utilisés, le référencement en leur sein est la première règle de l'accessibilité du document. À cette fin, des grandes structures centralisées ont la chance de pouvoir travailler sur leur référencement, et d'apporter de la visibilité à leur contenu. En ce sens, les grandes archives ouvertes bénéficient plus que des plus petites. En France, HAL profite d'un référencement plus que correct dans le moteur de recherche Google Scholar. Pour la dissémination d'articles scientifiques, il semble pertinent de verser dans l'archive ouverte pour gagner en visibilité. On note la même chose, dans la communauté du patrimoine : la centralité d'un site comme Gallica, bien représenté dans les moteurs de recherche, en fait un réservoir idéal pour les contenus que l'on souhaite exposer : dans le cas d'une bibliothèque numérique locale, il convient de penser un outil pratique et ergonomique, reconnaissable par les usager-e-s, ce qui pose des questions de coûts et de compétences. À une heure où des impératifs de réduction des coûts se multiplient, le versement de contenu dans une institution centrale, ou le passage par une solution Marque Blanche que nous avons étudié, fait sens.

Encore une fois, cela nuit à la décentralisation d'OAI-PMH, mais il faut comprendre que le système a très bien fonctionné aussi parce qu'il permettait à chaque institution d'exposer des

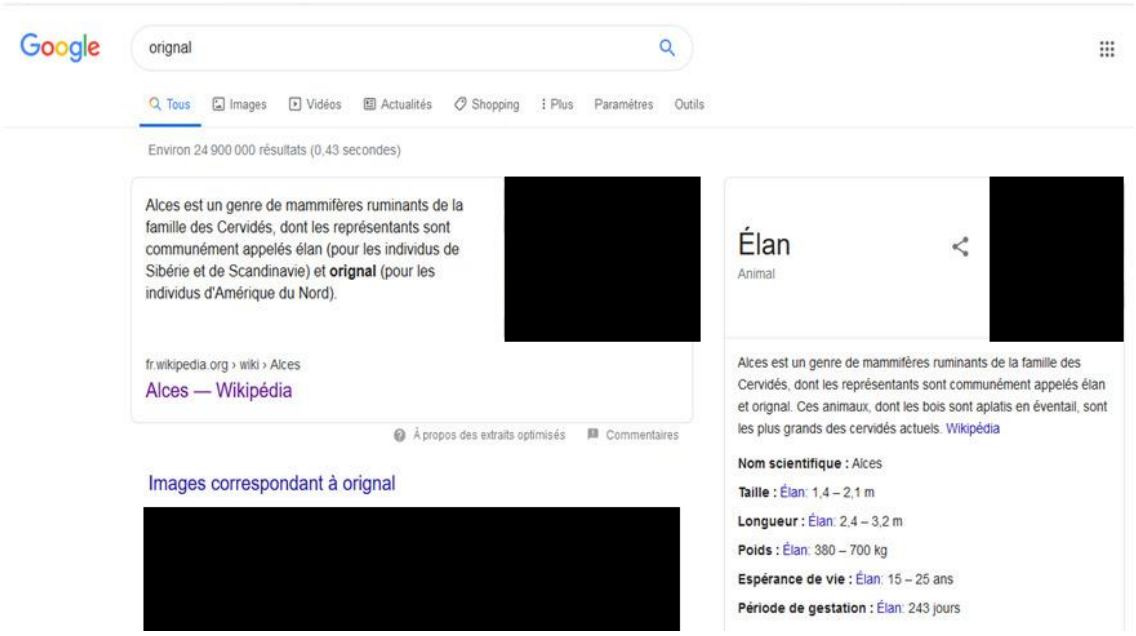
[94]GAUDINAT, Arnaud, BEAUSIRE, Jonas, FUSS, Megan, et al. Global picture of OAI-PMH repositories through the analysis of 6 key open archive meta-catalogs. *arXiv:1708.08669 [cs]* [en ligne]. Août 2017. [Consulté le 23 janvier 2020]. Disponible à l'adresse : http://arxiv.org/abs/1708.08669. ArXiv: 1708.08669.

métadonnées sur des documents tout en les conservant. Par exemple, le substitut numérique d'une thèse restait hébergé sur les serveurs de l'université, et ses métadonnées étaient exposées dans OAIster. En suivant le lien trouvé dans OAIster, il était possible d'accéder à la thèse sur le site de l'université. C'était donc un enjeu de visibilité pour les établissements, ce qui se voit également dans le domaine de la culture : OAI-PMH a permis de construire des projets et partenariats où les différents partenaires conservaient la maîtrise de leurs contenus. Aujourd'hui, la multiplication des contenus pose plusieurs questions importantes aux établissements, en terme de stockage, de conservation et de maintien. Est-il toujours aussi pertinent pour un musée de petite taille de conserver des collections numérisées importantes, sachant que donner l'accès à ces collections demande des serveurs, des équipes informatiques, pose la question de la pérennité des données, donc questionne le coût et la soutenabilité du projet ? Le web sémantique pose les mêmes questions en termes de centralisation et décentralisation : alors que son concept veut respecter la structure originelle du *World Wide Web*, la tendance actuelle est à la constitution de gigantesques bases de données, contenant toute l'information. Wikidata, qui fonctionne avec ses propres préfixes (wd, wdt, p), est cependant construit sur un modèle se rapprochant du Resource Description Framework, donc un standard du web sémantique ; mais c'est, par essence, une base extrêmement centralisée, et l'écosystème du web va dans le sens d'une centralisation encore plus accentuée. En effet, Google possédait sa propre base de données, Freebase, lancée en 2007, et en 2014, l'entreprise choisit d'arrêter le projet, et de verser la totalité du contenu dans Wikidata[95]. De cette manière, Wikidata devient beaucoup plus riche, et il est plus aisé d'accéder à cette information centralisée. L'enjeu du contrôle de l'information n'est plus le même, à l'heure actuelle, que dans les années 2000. Pour les acteurs majeurs du numérique, l'importance est l'accès à la donnée, et la centralisation est une solution de choix, notamment pour entraîner des machines à effectuer des tâches de manière automatisée (technologies de l'Intelligence Artificielle) ou pour améliorer le système de recherches sur le web (à l'exemple du *Knowledge Graph* de Google, sur lequel nous reviendrons).

Dans ce cadre, une information structurée est extrêmement intéressante, puisqu'elle est compréhensible par la machine. De cette manière, la prédiction de Berners-Lee, Lassila et Hendler se réalise en quelque sorte aujourd'hui, par exemple avec le *Knowledge Graph* de Google, qui crée un rectangle d'information qui apparaît à côté des résultats lors d'une

[95]Un contenu qui, par ailleurs, reste (en partie) à transférer. Sur le sujet, voir : Wikidata:WikiProject Freebase - Wikidata. Dans : *wikidata.org* [en ligne]. 17 décembre 2014. [Consulté le 11 février 2020]. Disponible à l'adresse : https://www.wikidata.org/wiki/Wikidata:WikiProject_Freebase.

recherche sur ce moteur ; les informations qui y apparaissent viennent de Wikidata, et leur présence est due à la compréhension par la machine de la nature de la recherche. Sur l'exemple ci-dessous, par exemple, on voit apparaître, dans ce rectangle, le terme « élan », puisque la machine trouve que le terme « orignal », dans Wikidata, est une variante (« également connu comme ») de l'élan, et rattache donc cette requête à un URI (ici, https://www.wikidata.org/wiki/Q35517). De cet URI viennent les informations qui sont indiquées dessous, à savoir « Nom scientifique », « taille », « longueur », « poids », « espérance de vie » et « période de gestation ».



Copie d'écran d'une recherche Google, images retirées

En soi, cet exemple rejoint et montre les limites de l'exemple que donnait Tim Berners-Lee : la machine a bien compris la requête, mais elle n'est pas allée chercher dans un grand nombre de bases de données, seulement dans cette gigantesque base de données centralisée. La réalisation paraît donc dérisoire au vu de la promesse, mais son potentiel de transformation n'en est pas moindre : en septembre 2019, une étude a montré que 49% des recherches sur Internet n'aboutissaient pas à un clic sur une autre ressource. Les raisons sont de deux types : non-pertinence des résultats proposés, qui impose une seconde recherche, ou bien l'inverse, c'est à dire des résultats trouvés sans même avoir besoin de cliquer sur un lien. La deuxième raison renvoie à l'implémentation du *Knowledge Graph* par Google : toutes les recherches sur le web ne sont pas approfondies, et si l'objet de la recherche est directement identifiable, pourquoi parcourir d'autres pages web ? Si le nombre de recherches sans clics continue d'augmenter, c'est la structure de la recherche d'information qui risque d'évoluer, l'enjeu

actuel étant un meilleur référencement, c'est-à-dire une position parmi les premières pages web retournées par Google à la suite d'une recherche. L'un des enjeux de la transition bibliographique, initiée par la BnF et l'ABES, est d'ailleurs de mieux faire remonter les collections des bibliothèques dans les moteurs de recherche, comme le dit Marianne Clatin (responsable du service Prospective et services documentaires de la BnF) en 2016 :

> « Il faut se mettre sur le chemin des usagers du web. Les usagers quand ils cherchent quelque chose, ils cherchent d'abord sur le web (tout comme nous!). Donc il faut être sur le web. Certes nos catalogues sont en ligne mais seule notre première page (notre site ou notre portail) est accessible à une recherche. Le contenu de nos catalogues se trouve lui dans le web profond. Il faut arriver à faire remonter ces données.»[96]

Sortir des références du web profond, voilà la mission que se donnent les bibliothécaires et les professionnel-le-s de l'information depuis l'apparition de cette technologie. OAI-PMH a été, à cet égard, un premier pas : en sortant l'information du silo catalogue, le protocole l'a placé dans l'environnement du web, avec le lot de problèmes que nous avons vu, et en le plaçant dans un nouveau silo, celui d'un entrepôt. Avec le programme de la Transition Bibliographique, c'est un nouvel effort qui apparaît pour mettre des ressources sur le chemin des usagers et usagères dans l'environnement web. Mais le changement technologique et la dissémination de cette technologie dans les institutions avancent dans deux temporalités très différentes : des entrepôts OAI-PMH s'ouvrent encore à l'heure actuelle, le protocole Z39-50 reste utilisé, et certaines bibliothèques s'informatisent pour la première fois en 2020. La stabilité et l'usage répandu sont donc les deux meilleurs atouts, aujourd'hui, d'un protocole certes imparfait. Il convient alors de repenser l'expression qui se situe dans le titre de ce travail : il n'y a peut-être pas d' « heure du web sémantique » de la manière dont on l'imaginait, et l'avènement qui semblait promis pour ces technologies n'adviendra pas dans la forme révolutionnaire, disruptive que l'on imaginait. Mais nous rentrons peut-être dans un moment où une partie des technologies, de la logique du web sémantique va infuser, et particulièrement dans les professions des sciences de l'information et des bibliothèques, par exemple avec le programme Transition Bibliographique[97], sur lequel nous aurons l'occasion de revenir.

[96]CLATIN, Marianne. Voyages vers « l'innovation en bibliothèque » 11. Dans : *Association des Bibliothécaires de France* [en ligne]. [s. d.]. [Consulté le 1 février 2020]. Disponible à l'adresse : http://www.abf.asso.fr/14/914/2006/ABF-Region/voyages-vers-l-innovation-en-bibliotheque-11.

[97] La transition bibliographique : des catalogues vers le web de données. Dans : *www.transition-bibliographique.fr* [en ligne]. [s. d.]. [Consulté le 14 février 2020]. Disponible à l'adresse : https://www.transition-bibliographique.fr/.

ÉCHANGER L'INFORMATION DANS L'ENVIRONNEMENT DES BIBLIOTHEQUES : ENTRE DEFIS ET PROMESSES

QUE FAIRE ?

Dans le contexte que nous venons d'étudier, il paraît difficile de se baser sur une tendance ressentie. À notre sens, les choix de technologie sont, en vérité, secondaires. En effet, OAI-PMH a été utile et est encore utile, puisqu'il a servi des objectifs humains, que se sont donnés les institutions et les structures. Plutôt que de se demander, si le web sémantique est sur le point d'advenir, si une autre solution qu'OAI-PMH existe, les professionnels de l'information se questionnent sur les objectifs desquels découleront une stratégie, et en bout de chaîne, des choix technologiques particuliers. En ce sens, on peut analyser l'apport d'OAI-PMH : il déplace des métadonnées depuis un entrepôt vers d'autres. On ne peut pas dire qu'il améliore réellement le référencement dans les moteurs de recherche, comme cela avait été espéré, ni qu'il favorise une interopérabilité globale, car même dans une implémentation minimaliste, conversions et adaptations sont nécessaires. Mais il fait circuler de manière efficace des métadonnées, et a permis de créer des services à valeur ajoutée sur le web, notamment la constitution de collections diverses dans des bibliothèques numériques. À un moment où le contexte a clairement changé, notamment vis-à-vis des problématiques de centralisation des données, d'adoption limitée des technologies du web sémantique, il importe de définir des objectifs, des priorités claires afin de faire, de la manière la plus commune possible, le choix de solutions technologiques compatibles et interopérables, ou tout du moins compatibles.

La question demeure : dans un monde où les changements structurels et l'adoption de standards généralisés se font sur le temps long, est-il utile, voire souhaitable, de viser l'interopérabilité maximale ? L'avenir n'est-il pas plutôt dans une généralisation de la conversion, c'est-à-dire à assumer l'hétérogénéité des contenus informationnels ? Si l'interopérabilité et la décentralisation ont été à la base du web, de son adoption et sa diffusion, aujourd'hui, en matière de contenus documentaires, l'interopérabilité générale est complexe à mettre en forme. Lars Svensson, dans son discours en ouverture des Journées ABES en 2014, déclarait ainsi que l'interopérabilité nécessitait « plus d'une seule

transformation des données »[98], mais aussi qu'en matière d'interopérabilité, le niveau qu'il importait de viser était « juste suffisant », et que ce niveau naissait des collaborations et échanges entre communautés. Au sein de ces communautés, des formats anciens, voire antiques d'un point de vue technologique, comme MARC, existent toujours, et restent utilisés : les faire évoluer est une entreprise qui demande du temps et une adaptation, pour aller vers cet objectif d'interopérabilité.

La mise au point d'une manière de décrire les documents sur le web, *Resource Description and Access* (RDA) cherche à créer un cadre général, mais se heurte aux particularismes. Au niveau français, la traduction progressive de RDA en un profil français, RDA-FR est aussi une adaptation de son contenu aux logiques propres de la France. La conception d'un format MARC nouvelle génération[99] marque également une particularité vis-à-vis d'autres formats actuellement en développement pour se mettre en conformité avec RDA, comme Bibframe (développé notamment par la Bibliothèque du Congrès). Ce dernier est en développement depuis près de 10 ans, et actuellement dans une phase de refonte[100], notamment pour englober la possibilité de décrire d'autres types de contenus.

Par cet exemple, on voit que l'interopérabilité totale, l'adoption d'un format unique, d'une logique commune, est incroyablement complexe, particulièrement dans des institutions dotées d'une longue histoire. De cette histoire naît une pratique, et de cette pratique naissent des besoins. Certains sont communs, d'autres sont propres à une institution, voire même à un département au sein d'une institution. Dans ce cadre, la mise en place d'initiatives communes risque de ressembler à ce qu'a été OAI-PMH : la mise en place d'un cadre lâche, dans lequel des institutions disposeront d'une marge de liberté. Dans ce contexte, l'interprétation du cadre rendra nécessaire, à nouveau, *mappings* et conversions pour permettre aux systèmes de discuter.

La limite atteinte ici n'est pas celle des machines, mais celle des personnes, qui construisent les systèmes d'information dans lesquels les machines ne sont que des outils. Les besoins particuliers, ou ressentis comme tels, sont ceux qui séparent les professions de

[98] "True interoperability needs more than one-way transformation", traduction personnelle de SVENSSON, Lars. *Jusqu'où l'interopérabilité est-elle nécessaire?* [en ligne]. Montpellier, 20 mai 2014. [Consulté le 20 février 2020]. Disponible à l'adresse : https://pdfslide.net/internet/jusquou-linteroperabilite-est-elle-necessaire.html.

[99] ROCHE, Mélanie et PEYRARD, Sébastien. *À défaut d'enterrement : les défis et les promesses de l'INTERMARC nouvelle génération*. 2018.

[100] Notamment à travers l'initiative Linked Data For Production ; voir par exemple la présentation suivante : ANDREW W. MELLON FOUNDATION et LIBRARY OF CONGRESS. *LD4P : Linked Data For Production* [en ligne]. 2018. [Consulté le 5 février 2020]. Disponible à l'adresse : https://www.loc.gov/bibframe/news/pdf/ld4p-alamw2018.pdf.

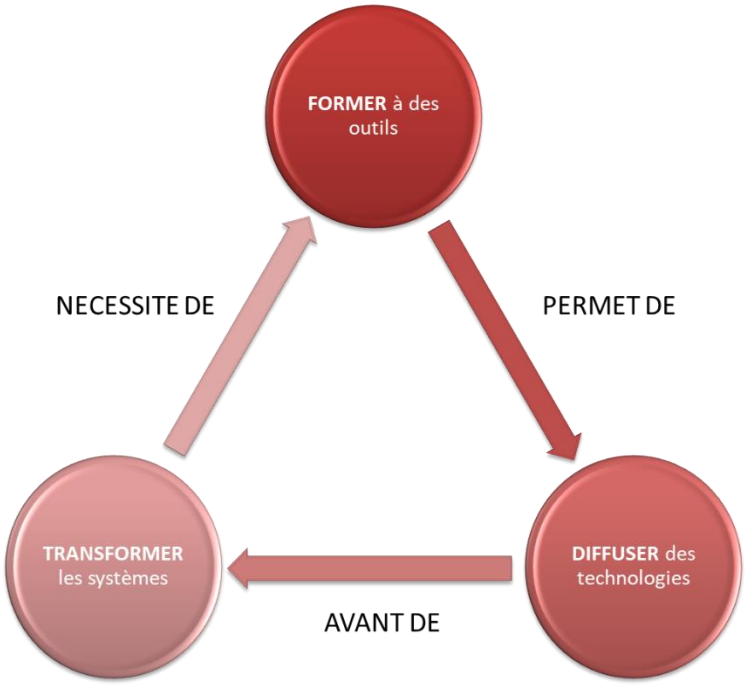
l'interopérabilité totale, qui permettrait aux machines comme aux structures humaines de dialoguer entre elles. Mais il convient de se demander si cette interopérabilité, si le niveau de granularité et de connexion que les professionnels de l'information souhaitent donner à l'information correspond réellement à un besoin. Les logiques de graphe, d'interconnexion du web sémantique semblent être un progrès, mais sont-elles une fin en soi ? Parce que les technologies du web sémantique ne sont pas aussi développées que celles du web social, les utilisateurs et utilisatrices n'y sont pas habitué-e-s : en 2017, Antoine Courtin, responsable de la cellule ingénierie documentaire à l'Institut National de l'Histoire de l'Art, rappelle que les pratiques des chercheurs, en terme de données, se limitent bien souvent à copier des données dans un document de traitement de texte[101]. Dans ce contexte, est-il opportun de chercher à adopter des formats, des langages de requête, en somme des logiques qui ne sont pas ceux des communautés que l'on sert ? Si l'on assume une vocation de formation des publics, il convient de pousser à des applications qui concrétisent les technologies du web sémantique, et changent alors radicalement, par exemple, la visualisation des données. La bibliothèque de Vaulx-en-Velin, par exemple, propose (en plus d'une vue plus classique) une visualisation des données en graphe[102]. Si, à l'inverse, on se place d'un point de vue où les pratiques des usager-e-s nous guident, on ne peut pas demander comme préalable la compréhension de logiques informatiques complexes, et les données que nous fournissons doivent être dans le format souhaité, même si celui-ci constitue une perte des enrichissements ou des liens qui ont été conçus. Ces deux points de vue, bien entendu, ne sont pas mutuellement exclusifs : on peut contribuer à amener des usagers, par exemple des chercheurs, vers des pratiques différentes, et par ce fait se constituer comme des évangélistes de technologies du web sémantiques (par exemple).

De notre point de vue, l'usage des technologies du web sémantique a un potentiel fort d'amélioration des services rendus aux usagers, mais de ces objectifs découlent des approches totalement différentes. Utiliser ces technologies en *back office* est une possibilité : répondre à des questions spécifiques d'usager-e-s avec SPARQL ou fournir des données par des extractions à la demande, par exemple, sans faire la promotion du moyen utilisé. Cela ne transforme pas les pratiques des usager-e-s, mais peut permettre de délivrer un service de

[101]POUPEAU, Gautier, COURTIN, Antoine, BERMÈS, Emmanuelle, et al. Quel renouvellement des formes de collaboration entre chercheurs et institutions patrimoniales ? Dans : *YouTube* [en ligne]. 14 octobre 2017. [Consulté le 1 février 2020]. Disponible à l'adresse : https://www.youtube.com/watch?v=WDpXvKTcgaQ.

[102]Voir par exemple cette notice : *The Beatles Anthology* [en ligne]. [s. d.]. [Consulté le 5 février 2020]. Disponible à l'adresse : https://bm.mairie-vaulxenvelin.fr/graphe?id=b2f59774-dff7-45ab-b089-ba971aa8cc7c##syrtis_search_page.

qualité. Utiliser ces technologies lors de projets particuliers, avec des publics intéressés, notamment pour des travaux de recherche, permet de diffuser une pratique, et de permettre à plus de monde d'exploiter nos données dans leur richesse : c'est un objectif de formation à des outils et de diffusion de technologies. Choisir d'aller plus loin, et d'utiliser ces technologies pour changer de manière générale les pratiques, c'est assumer un rôle de transformation. On peut concevoir ces trois points comme les parties d'un cycle plus large, qu'on peut définir ainsi :



Selon le diagnostic que l'on pose sur une situation, on définira donc des solutions différentes pour répondre à un besoin. Et trouver une solution qui permette d'aller vers l'interopérabilité nécessite d'atteindre un certain degré de consensus. Pour se mettre dans un point de vue prospectif, il importe donc de tracer des objectifs différents, qui aboutiront chacun à des solutions, des préconisations et des interrogations différentes. C'est ce que nous nous proposons de faire à présent, en ne cherchant pas à donner une solution monolithique, applicable en l'état et immédiatement. À la place, nous tenterons de brosser un tableau de solutions envisageables et de questionnements nécessaires, avec l'idée qu'il est important de prendre conscience des possibilités et des problématiques à l'œuvre afin de faire un choix informé, qui puisse aller dans la direction d'un consensus le plus large possible.

Ces objectifs doivent nous amener vers l'utilisation de solutions techniques, une seule ou plusieurs, combinées, augmentées ou séparées. Dans cette partie, nous nous attacherons donc à tenter de dégager plusieurs voies possibles, qui ne sont pas mutuellement exclusives, mais qui représentent des manières différentes d'appréhender les problématiques d'exposition et d'échange de données en bibliothèque et au-delà. Il peut s'agir de capitaliser sur l'existant, de s'appuyer sur des technologies certes imparfaites, mais très bien implantées, pour améliorer leur usage. Il peut être question de continuer un travail de diffusion des technologies du web sémantique, et de commencer à répandre leurs usages en partant des grandes institutions centralisées. Enfin, il est important d'évoquer aussi la possibilité, sinon la nécessité, de faire évoluer la formation et les usages des professionnel-le-s de l'information.

TRANSFORMER LES ARCHIVES INSTITUTIONNELLES

Dans le cadre de ce travail, nous avons abondamment signalé les défauts et les problèmes liés à l'utilisation du protocole OAI-PMH ; au moment de dresser un bilan et de chercher des perspectives, il convient toutefois de rappeler ses forces. La première est sa solidité : le protocole fonctionne, sans *bugs* majeurs, n'a pas besoin d'être entretenu ni remis à jour. La seconde est son maillage : dans les communautés professionnelles, les fournisseurs de solutions informatiques, il est répandu, et peut être mis en place de manière assez simple. La troisième est son asynchronie : grâce à cela, les moissonnages sont plus rapides et risquent moins d'erreurs si un entrepôt disparaît.

Pour optimiser l'échange de données en bibliothèque, il peut donc être intéressant de capitaliser sur ces forces du protocole OAI-PMH : plutôt que de chercher à le remplacer par une autre technologie, il peut être une manière de diffuser des bonnes pratiques en matière d'exposition et de partage de données. Il peut être une forme de cheval de Troie ouvrant la voie à des technologies plus riches.

Un cheval de Troie numérique ?

On a vu avec ORI-OAI qu'il était possible d'aller au-delà de l'implémentation minimale du Dublin Core. Si la solution décentralisée perd en utilisabilité, le fait d'améliorer la qualité des données exposées reste un objectif viable, et, de notre point de vue, particulièrement pertinent à un moment où les partenariats de gré à gré forment la quasi-

totalité des utilisations du protocole OAI-PMH. Dans ce cadre, des acteurs centraux, qui sont à la fois les principaux fournisseurs et les principaux exposants de données, ont la possibilité d'impulser des pratiques en termes de classification de l'information. Ce qui importe dans cette solution, c'est qu'augmenter la qualité des données échangées, c'est augmenter la qualité des services que l'on peut rendre après le moissonnage d'un entrepôt. C'est également un moyen de diffuser, dans le cadre d'une solution connue (OAI-PMH), des langages informatiques ou des structures différentes.

L'exemple de McGill : un OAI-PMH enrichi

L'université canadienne McGill, située à Montréal, en profitant d'un changement de fournisseur pour leur *institutional repository* (archive institutionnelle), a entrepris de travailler à une exposition riche des contenus[103]. C'est une manière de repenser le rôle de cette archive particulière qui est celle de l'institution. Si en France, une solution centralisée est en passe de devenir l'option principale (avec HAL), dans le reste du monde, les archives institutionnelles locales restent une composante importante de l'exposition de contenus scientifiques, en Europe et dans le reste du monde. À McGill, le principe était donc de profiter d'une migration de données pour repenser la manière de structurer ces données, qui doit s'insérer dans un critère d'interopérabilité. Puisque le protocole OAI-PMH reste, à l'heure actuelle, extrêmement utilisé, les données exposées doivent être moissonnables *via* ce protocole, mais l'interopérabilité n'est pas condamnée à se faire par le bas – c'est à dire par une exposition minimale de données en Dublin Core. Ainsi, en construisant neuf *work types* (type de travaux), chacun assorti d'une modélisation de données particulières, on aboutit à une exposition riche et structurée des contenus de l'archive institutionnelle. Mais l'intérêt de cet exemple est plutôt dans la construction de ces *work types*. De la même manière que dans la définition de l'European Data Model que nous avons analysé, ou encore dans le modèle de données de la BnF, l'intérêt est de construire des modèles appropriés aux usages. Ainsi, répliquer l'un de ces deux modèles serait absurde : les contenus gérés, la structure de l'institution, tous ces paramètres non informatiques doivent guider la création du modèle de données. Dans le cas de McGill, les neuf *work types* correspondent au contenu que l'institution propose, ou souhaiterait proposer, dans son archive institutionnelle, à savoir :

[103] DESMEULES, Robin Elizabeth. *Entretien sur la création d'un modèle de données compatible avec OAI-PMH au sein de l'université McGill*. 20 février 2020.

articles, thèses, rapports, communications en congrès, posters, livres et chapitres, images et critiques.

Afin de définir des profils pour chacun de ces types de documents, il est crucial de prendre en compte le modèle de données actuel, afin de juger de sa pertinence et des améliorations possibles. Pour certaines informations, Dublin Core remplit son office, et le changer compliquerait l'utilisation des données. Pour les informations qui échappent au modèle de données original, les équipes de McGill ont cherché à maintenir un équilibre entre des ontologies qui correspondaient à leur besoin et des ontologies qui étaient utilisées dans la communauté des bibliothèques et de la recherche.

Cet équilibre est extrêmement intéressant, et est valable dans les cas où l'on choisit d'étendre l'exposition des données : souvent, un standard global, mais imparfait, sera plus utile qu'un autre très peu adopté. C'est ce qu'on a vu avec Dublin Core, qui, malgré son aspect fruste (du moins dans la version simple), est aujourd'hui un terrain commun (malgré les différences d'implémentation). Suivre les pratiques de la communauté tout en travaillant à une meilleure exposition des données, c'est aller dans le sens d'une amélioration globale de la manière dont nos contenus sont visibles sur le web. En ce sens, OAI-PMH peut servir de vecteur pour propager de meilleures manières d'exposer nos données : le protocole est répandu, et si son obsolescence est souvent critiquée, il conviendrait de noter aussi que son adoption est une opportunité. Dans le contexte du Canada, par exemple, la mise à jour du standard de description des thèses *Electronic Theses and Dissertations*, ou ETD (qui est en partie la base de TEF en France[104]) a lancé la réflexion sur l'extension de la description des données académiques. L'université d'Alberta a été, dans ce domaine, une des pionnières au Canada. L'exemple de leur modèle de données, suivi en partie par McGill, permet la diffusion de standards d'exposition de données qui pourraient être découvertes et exposées par d'autres moyens que le protocole OAI-PMH. Si les exemples d'implémentation maximalistes d'OAI-PMH ne sont pas extrêmement nombreux, il reste possible de trouver des cas d'usages comme celui que l'on vient d'analyser. Il peut donc être intéressant, plutôt que de chercher à remplacer un protocole répandu, de capitaliser sur son adoption pour favoriser des implémentations plus riches, et en cherchant à créer des profils d'application qui puissent être réutilisés et augmentés par la communauté.

[104] « Pourtant, chacun reconnaît dans ces éléments des éléments issus du Dublin Core, du format bibliographique MARC 21 et du jeu de métadonnées ETD-MS », extrait de la recommandation : AFNOR. *Les métadonnées des thèses électroniques françaises-TEF* [en ligne]. mai 2005, p. 13. Disponible à l'adresse : http://www.abes.fr/abes/documents/tef/recommandation/tef_01.pdf.

Des possibilités pour le patrimoine et la lecture publique

L'exemple que l'on a vu concerne, au premier chef, les bibliothèques universitaires. Mais on peut constater d'autres initiatives qui utilisent également le protocole OAI-PMH, à l'heure actuelle, dans le domaine du patrimoine et de la lecture publique.

En 2017, un partenariat noué entre la BnF et la British Library pour mettre en valeur des manuscrits médiévaux[105] aboutit ainsi à l'utilisation de l'International Image Interoperability Framework[106] (IIIF) couplé avec le protocole OAI-PMH. Comme on l'a vu, le protocole est, historiquement, au cœur des partenariats de la BnF avec d'autres institutions. Mais depuis les premières collaborations, de nombreuses initiatives pour mieux exposer et échanger les données ont été engagées dans les bibliothèques à vocation patrimoniale. Parmi ces efforts, on peut citer IIIF, une initiative dans laquelle la BnF a été pionnière : c'est une solution technique qui permet notamment de visualiser des contenus distants tout en restant sur une page web. Dans l'exemple de la plateforme commune BnF-British Library, par exemple, le fait de sélectionner un document ne renvoie pas vers le site de l'une des deux institutions, mais permet à l'internaute de visualiser l'image, à travers un outil de visualisation, sur la plateforme elle-même. C'est une manière de servir le contenu à un internaute qui ne nécessite pas de duplication de l'image ni d'échanges complexes entre institutions. Le contenu distant est décrit dans un manifeste IIIF, dans un format JSON-LD (qui est un vocabulaire du web sémantique).

[105] SIRI, Francesco. France-Angleterre 700-1200 : manuscrits médiévaux de la Bibliothèque nationale de France et de la British Library, un programme de la Fondation Polonsky. Dans : *MANUSCRIPTA : Manuscrits médiévaux conservés à la BnF* [en ligne]. [s. d.]. [Consulté le 20 février 2020]. Disponible à l'adresse : https://manuscripta.hypotheses.org/241.

[106] *Home — IIIF | International Image Interoperability Framework* [en ligne]. [s. d.]. [Consulté le 28 juin 2019]. Disponible à l'adresse : https://iiif.io/.

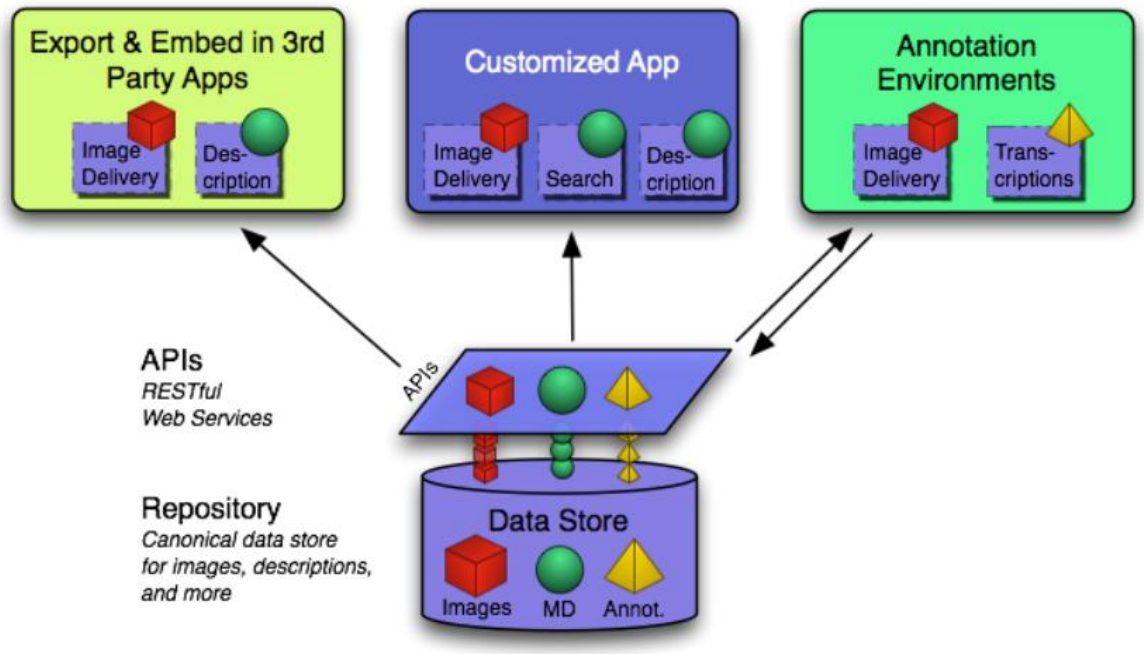


Schéma de fonctionnement de IIIF (pris dans la présentation de Tom Cramer[107])

La mise en place de la plateforme commune, toutefois, s'est faite *via* le protocole OAI-PMH. Mais celui-ci, en plus des métadonnées descriptives habituelles en Dublin Core, a permis de transmettre le manifeste IIIF des documents agrégés sur la plateforme. De cette manière, chaque institution conserve ses documents, expose des métadonnées parmi lesquelles le manifeste qui permettra aux internautes de visualiser le contenu sur la plateforme tierce. Encore une fois, on voit ici qu'OAI-PMH sert de porteur de technologies ou de pratiques plus innovantes, qui ont un potentiel transformatif dans nos communautés professionnelles. Il est, jusqu'à un certain point, adaptatif, puisqu'il permet des usages plus larges que ce à quoi ressemble son implémentation dans la plupart des cas à l'heure actuelle. Mais, si cette expérimentation est étendue à de nombreux partenariats, la technologie IIIF se répandra dans les différentes communautés professionnelles. De cette manière, on peut concevoir le protocole comme un outil, un levier pour transformer les pratiques des bibliothèques, par ce qu'il peut transporter.

D'autres expérimentations s'appuient sur le protocole : l'application Bibliotouch[108], développée par l'École Nationale Supérieure des Sciences de l'Information et des Bibliothèques (ENSSIB) et l'entreprise Biin permet par exemple d'exposer les contenus de

[107] CRAMER, Tom. *Intro to IIIF* [en ligne]. Rijksmuseum, 18 octobre 2016. [Consulté le 27 février 2020]. Disponible à l'adresse : https://docs.google.com/presentation/d/133rKIAvHCixG22kDkCieqI_FCvklfv5maOZ3D3ueTeE/present?usp=embed_facebook.

[108] *Bibliotouch* [en ligne]. [s. d.]. [Consulté le 20 février 2020]. Disponible à l'adresse : https://bibliotouch.enssib.fr/#/.

son catalogue à travers une interface numérique pour faciliter la navigation et la découverte. Pour ce faire, elle moissonne le contenu de l'entrepôt OAI-PMH de la bibliothèque. Dans le cas de l'ENSSIB, l'entrepôt OAI-PMH est créé via un *plugin* du Système Intégré de Gestion de Bibliothèque (SIGB) Koha. Ce seront les métadonnées moissonnées qui permettront d'afficher les différentes publications dans l'application. Dans ce cas, le protocole sert de base à une nouvelle manière d'exposer des données, on est donc toujours dans l'idée de service à valeur ajoutée qui était défendue dans le rôle du fournisseur de service (on notera qu'ici, le fournisseur de service est également fournisseur de données.), mais dans un environnement qui n'a plus guère à voir avec le contexte d'énonciation du protocole.

Les exemples que nous venons de présenter montrent donc des voies possibles d'utilisation d'OAI-PMH au-delà d'une implémentation minimale : il peut être un outil facilitant la diffusion de bonnes pratiques, de nouveaux standards et de pratiques innovantes. Il s'agit en quelque sorte d'utiliser le protocole pour aller vers son dépassement.

Vers un réseau structuré et transformé

Comme nous l'avons vu avec l'exemple de Van de Velde, les concepteurs originaux d'OAI-PMH et des entrepôts institutionnels sont eux-mêmes assez critiques sur les performances et les fonctionnalités limitées qui en résultent aujourd'hui. À ces acteurs, il convient de rajouter les groupes qui gèrent des archives ouvertes. En effet, si ces dernières n'ont pas répondu totalement à la promesse originale de transformer la communication scientifique, leur apparition a été accompagnée d'une véritable mise en réseau, qui se fait notamment au moyen de la Confederation of Open Access Repositories[109] (COAR), qui a été lancée en 2009, à la suite d'une initiative qui rassemblait les archives ouvertes européennes. En mettant en réseau les différentes archives ouvertes, cette confédération permet également de réfléchir aux enjeux d'évolution de celles-ci. Pour répondre à ce mandat, COAR a publié en 2016 le résultat d'un travail d'enquête et de réflexion sur le futur des *repositories*, en s'appuyant sur les besoins des personnes qui les utilisent, les limites des briques informatiques et technologiques qui les composent et les possibilités offertes par d'autres technologies.

[109] ADMIN. About COAR. Dans : *COAR* [en ligne]. [s. d.]. [Consulté le 19 février 2020]. Disponible à l'adresse : https://www.coar-repositories.org/about-coar/.

Il s'agit, à travers le développement d'une nouvelle génération d'archives ouvertes, de répondre au même besoin que celui qui était reconnu en 1999 : la transformation de la communication scientifique. Le plan de COAR[110] est donc ambitieux, notamment dans les fonctionnalités qu'il reconnaît comme essentielles, avec par exemple de la fouille de texte et de données (*text and data mining*, ou TDM), des composantes sociales ou encore un processus de revue par les pairs (*peer-review*). Dans un moment où la confiance dans le système de l'édition scientifique est plus qu'érodée, et où les gouvernements européens et mondiaux s'emparent de la question de l'*open access*, nous nous situons donc dans un moment où l'archive ouverte peut devenir un levier d'action. Si des possibilités de *peer-reviewing* apparaissent dans des archives ouvertes, nationales, disciplinaires ou institutionnelles, et qu'elles sont utilisées massivement par la communauté de la recherche, il y a alors une possibilité de sortir du circuit de publication et de validation classique. Des composantes sociales (identification et authentification, commentaires et annotations de contenus) permettent d'intégrer, dans les archives ouvertes, une partie des fonctionnalités, les composantes sociales que les chercheurs et chercheuses apprécient dans ResearchGate ou Academia. Ce plan, en définitive, prend en compte la majeure partie des problématiques et limites que nous avons relevées dans ce travail. Pour ce faire, il s'appuie sur un ensemble de technologies, notamment Signposting et ResourceSync. Ces deux technologies, développées par Herbert Van de Sompel et Michael L. Nelson, déjà à l'oeuvre dans le développement du protocole OAI-PMH ou de la spécification OAI-ORE, visent à actualiser ces technologies. ResourceSync[111], basé sur le protocole Sitemaps, permet de moissonner du contenu distant, mais aussi des métadonnées.

[110] CONFEDERATION OF OPEN ACCESS REPOSITORIES. *Next Generation Repositories* [en ligne]. 15:50:55 UTC. [Consulté le 10 février 2020]. Disponible à l'adresse : https://www.slideshare.net/ukcorr/next-generation-repositories-115010494.

[111] *ResourceSync Framework Specification* [en ligne]. [s. d.]. [Consulté le 22 février 2020]. Disponible à l'adresse : http://www.openarchives.org/rs/1.1/resourcesync.

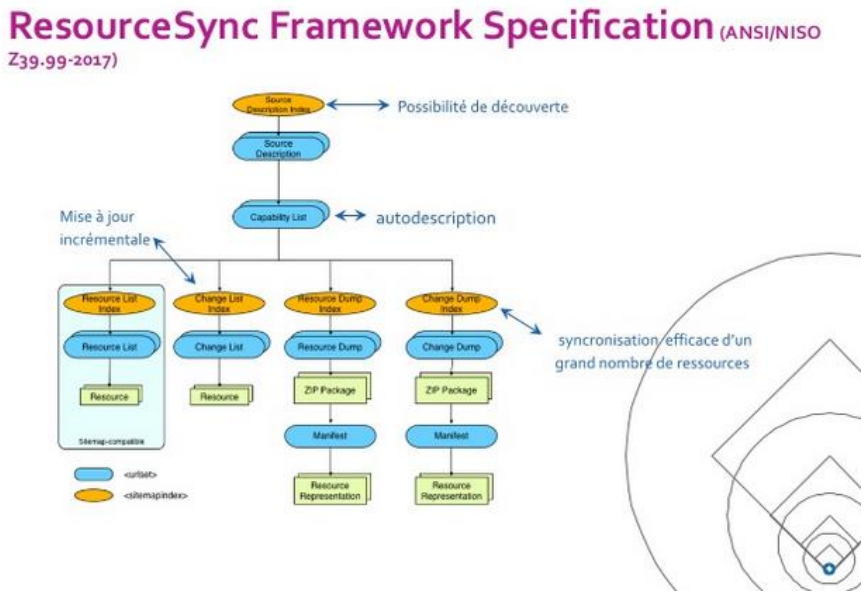


Schéma de fonctionnement de ResourceSync (extrait de la présentation de Pascal Aventurier[112])

Il correspond donc à ce changement de paradigme où l'on passe d'un intérêt envers les métadonnées à un intérêt envers les ressources elles-mêmes. ResourceSync permet de synchroniser les contenus, et ainsi de continuer à les fournir. Si le contenu est modifié, les changements sont répercutés, et s'il est supprimé, ResourceSync permet de le supprimer dans tous les endroits où il est exposé, afin de ne pas avoir, comme on pourrait en avoir *via* OAI-PMH, des métadonnées accompagnées d'un URL qui n'est rien d'autre qu'un lien mort. Signposting[113], lui, permet de décrire, à travers des technologies proches du web sémantique, les liens internes aux archives ouvertes, afin que les machines puissent comprendre l'information et la fouiller efficacement. De cette manière, un article peut être efficacement lié à un Digital Object Identifier (DOI), son identifiant, à l'ORCID de son auteur-ice ou encore les métadonnées bibliographiques qui permettent de citer l'article. C'est une autre manière de recomposer des contenus complexes, comme pouvait le faire OAI-ORE. Le choix de COAR de s'appuyer principalement sur ces deux technologies montre l'espoir qu'elles puissent essaimer, ainsi que l'avait fait OAI-PMH. Elles sont toutefois, toutes les deux, basées sur des standards du web sémantique, et donc plus adaptées à un environnement actuel. Utilisées en conjonction avec d'autres technologies, elles pourraient permettre, par

[112] AVENTURIER, Pascal. *Mettre en pratique les recommandations sur les archives ouvertes de n…* [en ligne]. 23:26:59 UTC. [Consulté le 27 février 2020]. Disponible à l'adresse : https://www.slideshare.net/paventurier/mettre-en-pratique-les-recommandations-sur-les-archives-ouvertes-de-nouvelle-gnration-de-coar-pour-larchive-ouverte-institutionnelle-horizon-pleins-textes-de-lird. Library Catalog: SlideShare.

[113] *Signposting the Scholarly Web* [en ligne]. [s. d.]. [Consulté le 22 février 2020]. Disponible à l'adresse : https://signposting.org/.

exemple, le *peer reviewing*, à travers des modifications qui sont synchronisées et assorties de notifications pour les acteurs impliqués dans le processus. Si ResourceSync et Signposting sont reconnues comme les deux briques informatiques qui permettent de rendre la plupart des services à valeur ajoutée, il convient cependant de voir que le plan de COAR se fonde sur un ensemble de technologies, comme on peut le voir sur la liste proposée sur le site[114]. Plutôt que de proposer une implémentation coordonnée de toutes ces technologies, le groupe de travail de ces archives de nouvelle génération suggère un ensemble qui, potentiellement, peut rendre les services jugés nécessaires. Il convient ensuite de diffuser ces technologies au sein des différentes archives, mais aussi et surtout auprès des fournisseurs de solutions logicielles (DSpace, Fedora, Eprints, Omeka...). De cette manière, à travers un mouvement de la base, il semble possible de transformer en profondeur, mais de manière progressive, les archives que l'on connaît. Il est toutefois intéressant de noter que, dans ce choix des archives de nouvelle génération, le protocole OAI-PMH est purement et simplement abandonné, au profit de ResourceSync. Si le plan proposé par COAR est suivi par la communauté, on pourrait imaginer, à moyen ou long terme, une quasi-disparition du protocole : les solutions alternatives, bien que techniques, seraient adoptables dans le cadre d'un plus petit nombre d'archives ouvertes, plus centralisées (nationales, régionales ou disciplinaires). Leur centralisation serait une plus-value dans un mouvement allant vers une interaction forte de la communauté des chercheurs, puisqu'elles bénéficient d'une visibilité importante. Dans le monde de la physique, des mathématiques et d'autres disciplines, arXiv représente par exemple une structure bien connue. En France, HAL est bien implantée dans les universités, et reconnue par le Ministère comme un acteur essentiel de la science ouverte. En transformant l'archive ouverte telle qu'on l'appréhende à l'heure actuelle, il serait possible de changer son rôle dans l'écosystème.

[114] *COAR Next Generation Repositories: Technologies* [en ligne]. [s. d.]. [Consulté le 19 février 2020]. Disponible à l'adresse : https://ngr.coar-repositories.org/technology/.

EXPLOITER LE WEB SEMANTIQUE CENTRALISE

Une autre possibilité, pour les bibliothèques, est de continuer un travail de sensibilisation et d'évangélisation des technologies du web sémantique elles-mêmes. Pour ce faire, en France, les deux opérateurs nationaux que sont l'ABES et la BnF ont un rôle central à jouer, en particulier via le programme de la Transition bibliographique.

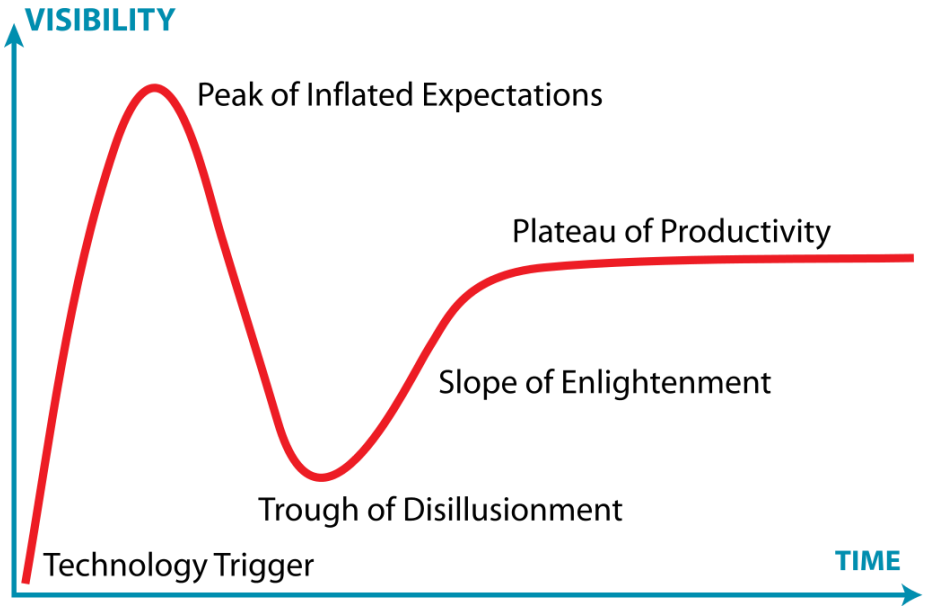
En transition vers un web de données

Nous avons auparavant évoqué la question des silos que constituent nos catalogues, une organisation qui a poussé des professionnel-le-s à promouvoir une libération de ces données, dans la logique du web sémantique. La Transition bibliographique est la traduction nationale de cette ambition, et elle s'applique en même temps aux bibliothèques universitaires (à travers l'ABES) et de lecture publique (à travers la BnF). La publication du code de catalogage Resource Description and Access (RDA) en 2010 amorce une réflexion française autour de son éventuelle adoption. Elle se déroule en plusieurs étapes[115] : tests par des groupes d'expert-e-s, mise en place d'un groupe stratégique, puis d'un comité stratégique bibliographique. Les différents acteurs de ces étapes incluent des bibliothécaires, mais aussi l'Agence Française de Normalisation (AFNOR), les différents ministères de tutelle et les organismes de formation. La variété de ces acteurs souligne bien les problématiques que nous avons tenté de mettre en évidence dans ce travail : le changement commun des pratiques implique une organisation englobante, et une progression prudente. Ne pas impliquer les différentes parties prenantes dans le processus, c'est prendre le risque de ne pas parvenir à créer un véritable changement, et donc perpétuer une incohérence forte dans l'exposition de nos données, ce qui pose un véritable problème pour répondre aux défis communs et particuliers qui sont ceux de la profession.

Comme nous avons essayé de le démontrer, le paysage du web n'est pas, à l'heure actuelle, celui qu'il était en 2010 : l'enthousiasme autour du web sémantique est très majoritairement retombé, et la perspective d'une transformation révolutionnaire n'est plus guère citée. Mais, pour autant, il n'est pas inintéressant de poursuivre et de mener à bien le programme, lancé officiellement en 2015 ; en effet, si l'on applique au web sémantique le

[115] Historique des travaux français sur RDA. Dans : *www.transition-bibliographique.fr* [en ligne]. [s. d.]. [Consulté le 18 février 2020]. Disponible à l'adresse : https://www.transition-bibliographique.fr/enjeux/historique-travaux-francais-rda/.

filtre du *hype cycle* (cycle de battage médiatique), on pourrait placer les technologies du web sémantique dans une des deux dernières catégories, c'est-à-dire la pente de l'illumination (*slope of enlightment*) ou le plateau de productivité (*plateau of productivity*).



Par Jeremykemp sur Wikipédia anglais, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=10547051

On peut situer le pic des attentes démesurées (*peak of inflated expectations*) dans le début des années 2010, alors que le gouffre de la désillusion (*trough of disillusionment*) serait vers 2016-2017. Actuellement, on a vu avec l'exemple du *Knowledge Graph* que si l'adoption est limitée, elle n'en transforme pas moins les usages, même de manière difficilement perceptible à l'heure actuelle. Nous nous plaçons donc dans un moment où des réalisations concrètes sont observables, même si ces technologies restent dans une sorte de niche du web. Plutôt que de s'enthousiasmer sur une autre technologie, et de recommencer le cycle ci-dessus, il semble pertinent de continuer à avancer dans celui qui est déjà entamé, dans le but d'atteindre une productivité, certes limitée par rapport à des espoirs initiaux, mais qui permettra tout de même une vraie plus-value dans le service rendu à des utilisateurs.

De la même manière qu'OAI-PMH, la Transition bibliographique avait pour objectif de faire mieux remonter les données des bibliothèques dans les moteurs de recherche, par des moyens différents. L'exposition de nos catalogues en données liées le permet, dans une certaine mesure : en recherchant un roman de Madeleine de Scudéry[116], publié entre 1654 et

[116] La recherche est la suivante : clélie histoire romaine - Recherche Google. Dans : *Google* [en ligne]. 18 février 2020. [Consulté le 18 février 2020]. Disponible à l'adresse : https://www.google.com/search?sxsrf=ACYBGNTw0BsgfjisX4iwnJCiY_WHRYvTcQ%3A1582020839895&ei=57hLXtKZNp6SjLsP

1660, on trouve data.bnf.fr et Gallica en septième et huitième position dans la première page de résultats du moteur de recherche Google. Mais cette remontée reste à relativiser : tout d'abord elle s'observe sur des titres et des auteur-ice-s relativement peu connu-e-s. Dans le cas d'auteur-ice-s plus célèbres, la remontée des données est relativement fluctuante, donnant, à l'occasion, de belles surprises, comme de trouver Gallica comme premier lien donné par Google, et parfois de plus mauvaises, c'est-à-dire être présent dans la troisième, voire la sixième page de résultats. Parfois, cette présence dans les premiers liens, ou dans la première page, évolue même - positivement ou négativement - dans le temps. De plus, comme on l'a vu, nombre de recherches, à l'heure actuelle, ne débouchent sur aucun clic dans la liste de sites. À ce titre, l'endroit où il conviendrait d'être présent serait dans le *Knowledge Graph* lui-même, mais celui-ci est constitué de références Google Books, et de Wikipédia. C'est ce point qui pose problème quand on vise un meilleur référencement : les moteurs de recherche sont seuls maîtres de leur algorithme quant à la sélection et à la présentation des résultats. Des protocoles comme Sitemaps ne sont accompagnés d'aucune garantie en terme d'indexation ; nous pouvons faciliter le travail de recherche des robots, mais rien ne nous garantit que les moteurs de recherche ne changeront pas demain ces protocoles, ou les manières de présenter les résultats. À l'heure actuelle, il paraît logique, si l'on prend la perspective de Google, de présenter ses propres ressources numérisées (*via* Google Books) plutôt que celles de Gallica. Mais le fait de faire remonter, à l'heure actuelle, des données de bibliothèque pour des auteurs et autrices peu connu-e-s, est en soi un succès. Cela correspond aux pratiques de recherche de certain-e-s internautes qui, elles et eux, cliquent sur une des ressources proposées, sont peut-être en recherche d'un autre site que Wikipédia, et pour qui la BnF peut représenter une véritable plus-value. Car, si les recherches sans clic représentent une tendance, il ne faut pas oublier la part des internautes qui utilisent différemment ces moteurs, qui sont à la recherche d'information spécialisée, approfondie ; ce peuvent être des chercheur-se-s, des professionnel-le-s du livre, des étudiant-e-s, qui font tous et toutes partie du public potentiel des bibliothèques. Si demain, Google transforme radicalement son algorithme, la présentation des pages de résultats, rien ne nous assure d'être encore pertinent-e-s sur ce point, mais, comme on va le voir, le programme de la Transition bibliographique comporte bien d'autres avantages ; tant que nous restons dans cet environnement de recherche, nous améliorons le service rendu aux usager-e-s potentiel-le-s.

y46n0Ag&q=cl%C3%A9lie+histoire+romaine&oq=cl%C3%A9lie&gs_l=psy-ab.3.3.0l3j0i67j0l6.20939.21968..23945...0.2..0.322.1145.0j4j1j1.....0....1..gws-wiz.....0i71j35i39j0i131.HuwlInMdCdE.

Répandre l'utilisation de SPARQL

L'exposition des données des bibliothèques en RDF rend possible leur interrogation *via* le langage de requête SPARQL, que nous avons expliqué un peu plus tôt. À ce point de vue, les promesses que nous avons évoquées ne sont pas exagérées : l'interrogation est possible, et propose de nombreuses utilisations, autant pour les professionnel-le-s que pour les utilisateur-ice-s, de manière visible ou plus discrète.

Pour comprendre les possibilités offertes, mais aussi les étapes à franchir, nous pouvons utiliser un exemple qui a eu un certain retentissement dans la presse. En décembre 2019, le journal *Le Monde* souhaite organiser un « palmarès des 101 romans des lecteurs » du journal[117], et, pour s'appuyer sur une base d'ouvrages qui permette de faire un choix large, demande à la BnF si elle dispose d'une pareille information, mais, selon *Le Monde,* cela est impossible car la fiction n'est pas indexée. Il est donc difficile, sinon impossible, d'extraire les documents de la manière dont le journal aurait souhaité. Le journal s'est donc tourné vers le site SensCritique, qui a donné accès à sa base de romans. En imaginant un alignement de toutes les données de la BnF avec des référentiels extérieurs (notamment Wikidata), il aurait sans doute été possible, à l'aide d'une requête croisée, de répondre. Mais la demande du *Monde* est intéressante à plus d'un titre : elle démontre une utilisation possible des données que possèdent nos établissements. En tant qu'établissement responsable du dépôt légal, la BnF est identifiée comme une ressource ; de plus, ses données sont produites par des agents publics, elles tombent donc dans la logique de l'*open data*.

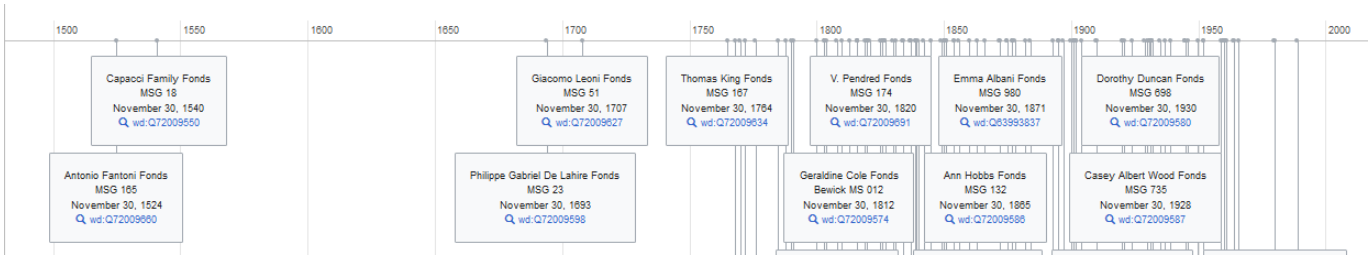
Les possibilités offertes au grand public ou aux organismes qui le visent (notamment informatifs, comme *Le Monde*) sont également présentes pour le monde de la recherche. Le projet CORPUS[118], actuellement dans un processus de transformation pour devenir bientôt le Data Lab de la BnF, est une initiative pour donner aux chercheur-se-s, en plus de l'accès aux documents, des outils pour les exploiter, les fouiller et les placer dans un contexte plus large. Dans ce cadre, les équipes de recherche peuvent être présentées aux équipes qui gèrent data.bnf.fr afin de profiter, notamment, du langage de requête SPARQL qui peut être utile pour définir un corpus particulier (par exemple, la totalité des ouvrages ayant un éditeur

[117] PARIENTÉ, Jonathan et FERRER, Maxime. 26 000 votants, 70 000 livres, 5 choix : comment le palmarès des 101 romans des lecteurs est né. *Le Monde.fr* [en ligne]. 27 décembre 2019. [Consulté le 18 février 2020]. Disponible à l'adresse : https://www.lemonde.fr/culture/article/2019/12/27/26-000-votants-70-000-livres-5-choix-comment-le-palmares-des-101-romans-des-lecteurs-est-ne_6024215_3246.html.

[118] BOUCHARD, Ariane. Présentation du projet CORPUS à la BnF. Dans : *Web Corpora* [en ligne]. [s. d.]. [Consulté le 18 février 2020]. Disponible à l'adresse : https://webcorpora.hypotheses.org/119.

particulier, toute la production sortie une année donnée, toutes les éditions d'un ouvrage...). Par ce genre d'initiative, on offre un service nouveau, étendu à la communauté de la recherche, ce qui est permis par l'exposition de données plus riches. Cela étend l'utilisation des collections de la BnF, mais ne se limite pas à cet établissement.

La diffusion des technologies du web sémantique peut ainsi se faire par le lien particulier tissé avec les chercheurs et chercheuses, mais aussi des personnels, par exemple pour organiser une exposition, créer une bibliographie particulière, effectuer une recension d'ouvrages, autant de cas où une requête SPARQL peut être utile. Diffuser ces technologies auprès d'un public amené à utiliser des données, à utiliser nos collections permet d'encourager des utilisations plus riches de celles-ci. Il faut toutefois préciser que ces utilisations, surtout auprès des chercheurs, se font souvent au cas par cas, et qu'il ne rentre pas dans les missions de la BnF, pas plus que dans celles de l'ABES, de former les utilisateur-ice-s au requêtage SPARQL. Mais reste la possibilité de former des personnes intéressées chez les professionnel-le-s et chez les usager-e-s, ou de s'appuyer uniquement sur les premier-e-s pour rendre des services à valeur ajoutée aux second-e-s. À la possibilité de retourner des listes de documents correspondant aux critères particuliers s'ajoutent aussi d'autres possibilités. Nous avons vu plus tôt que SPARQL permettait de trouver des dates, des coordonnées géographiques : le requêteur proposé par Wikidata permet de représenter ces données sous forme de cartes, de graphe, de frise chronologique.



Exemple de frise chronologique en réponse d'une requête SPARQL[119]

[119] Voir la requête correspondante et ses résultats sur : *Wikidata Query Service* [en ligne]. [s. d.]. [Consulté le 27 février 2020]. Disponible à l'adresse : https://query.wikidata.org/embed.html#%23defaultView%3ATimeline%0A%0ASELECT%20%3Finstance_of%20%3Fstart_time%20%3Ftitle%20%3Finventory_number%20WHERE%20%7B%0A%0ASERVICE%20wikibase%3Alabel%20%7B%20bd%3AserviceParam%20wikibase%3Alanguage%20%22%5BAUTO_LANGUAGE%5D%2Cen%22.%20%7D%0A%0A%3Finstance_of%20wdt%3AP31%20wd%3AQ3052382%3B%0A%0Awdt%3AP195%20wd%3AQ62535993.%0A%0A%3Finstance_of%20wdt%3AP580%20%3Fstart_time.%0A%3Finstance_of%20wdt%3AP1476%20%3Ftitle.%0A%3Finstance_of%20wdt%3AP217%20%3Finventory_number.%0A%0A%7D.

. Si celui de la BnF[120] comme celui de l'ABES[121] ne supportent pas cette fonctionnalité, il est tout à fait possible que des développements ultérieurs proposent ces options. On pourrait ainsi répondre, directement, *via* les données et les outils de nos institutions, à encore plus de demandes fines et précises des utilisateurs, et notamment les chercheurs.

Comme nous l'avons vu, le web sémantique n'a pas révolutionné de manière tangible les pratiques des utilisateurs et utilisatrices ; mais dans le contexte actuel, et en continuant une diffusion de ses technologies au sein de la communauté professionnelle des bibliothèques et de la documentation. Ainsi, elle se place en tant qu'adjuvante précieuse auprès de certains publics en recherche d'informations complexes. Les collections, particulièrement au sein de réservoirs nationaux, deviennent massives, hybrides : les professionnel-le-s de l'information doivent permettre à leurs publics de naviguer dans les collections, sans avoir à comprendre la complexité du fonctionnement des systèmes d'information, sans devoir faire l'apprentissage de langages informatiques et de systèmes de requêtage qui ne sont pas une partie de leur pratique habituelle du web. C'est, à l'heure actuelle, un véritable défi, pour les bibliothécaires notamment, de retrouver une place en tant qu'intermédiaire entre les publics et l'information souhaitée. L'autonomie des publics a été au cœur d'un certain moment bibliothéconomique. C'est d'ailleurs toujours le cas pour une certaine partie des bibliothèques de lecture publique : la recherche d'un roman de la rentrée littéraire, d'une bande dessinée tout juste sortie, d'un documentaire approprié sur un sujet doivent pouvoir se faire de manière simple et intuitive. Mais pour une recherche plus spécifique et exigeante, l'accompagnement des utilisateur-ice-s de nos collections doit devenir un impératif.

Avec cet accompagnement, la bibliothèque reste pertinente, la richesse de ses collections reste une véritable valeur ajoutée pour ses publics. Sans cette médiation, nos collections risquent de devenir de moins en moins utiles, utilisables et désirables. Dans le cadre de la Transition bibliographique, la montée en compétence doit se faire en réseau : si, au sein de la BnF, le travail avec des chercheurs, l'exploitation des données est déjà en cours, ce n'est pas nécessairement vrai dans toutes les bibliothèques municipales de France, qui disposent pourtant de collections vastes et protéiformes. Il est crucial de pouvoir signaler ces collections, de pouvoir accompagner leur exploitation par des publics, experts ou non. Le rôle des professionnel-le-s de l'information est à observer en miroir de celui qu'il a été dans

[120] *Virtuoso SPARQL query Editor* [en ligne]. [s. d.]. [Consulté le 22 février 2020]. Disponible à l'adresse : https://data.bnf.fr/current/sparql.html.

[121] *ABES : démonstrateur SPARQL* [en ligne]. [s. d.]. [Consulté le 22 février 2020]. Disponible à l'adresse : https://lod.abes.fr/sparql.

la fin des années 90, au moment de la naissance d'OAI-PMH : structurer une masse d'informations, et l'utiliser pour rendre des services ajoutés à une communauté identifiée. De la même manière, demain, l'utilisation de technologies du web sémantique pourra prendre plusieurs formes selon la typologie des établissements, et les ressources et données dont ils disposent. On peut d'ores et déjà imaginer des utilisations : une des trois bibliothèques numériques du Sillon Lorrain, Limédia Galeries, propose un document recensant « plus de 130 dessins d'édifices situés en Lorraine »[122]. Si ces cent-trente édifices sont liés à des coordonnées géographiques exprimées en RDF, il sera possible, avec un requêteur SPARQL adapté, de générer une carte montrant toutes les localisations des bâtiments recensés, dans la logique des *mashups*[123]. Cela peut intéresser aussi bien des personnes curieuses de l'histoire de leur région, de leur ville ou village que des érudits locaux et même des chercheurs et chercheuses travaillant sur l'histoire régionale. Dans cette logique, un système centralisé est préférable : un point d'interrogation unique facilite la recherche, la mutualisation des coûts reste avantageuse, et il n'est pas nécessaire pour une personne cherchant à créer un service à valeur ajoutée de connaître les différentes bases de données sur lesquelles créer une requête combinée. Mais de la même manière que la diffusion des technologies du web sémantique, le mouvement de centralisation que nous avons décrit est une tendance qui s'observe sur plusieurs années. On peut donc imaginer une période intermédiaire, voire même, à terme, le maintien d'une forme de décentralisation, avec des utilisateur-ice-s expert-e-s des données, qui connaissent le réseau des points d'interrogation existants. Ces personnes pourraient donc maîtriser des interrogations SPARQL croisées pour créer des services adaptés, qui aient, comme dans la logique d'OAI-PMH, une valeur ajoutée.

FAIRE EVOLUER LE POSITIONNEMENT DES PERSONNELS

Enfin, il est important, de notre point de vue, d'aborder la question du rôle des personnels, des structures et institutions dans les problématiques des données, de leur exposition, de leur échange, peut-être plus globalement dans les problématiques du numérique qui sont

[122] Table des planches du « Recueil de belles maisons, hôtels, châteaux exécutés en Lorraine ». Dans : *galeries.limedia.fr* [en ligne]. [s. d.]. [Consulté le 19 février 2020]. Disponible à l'adresse : https://galeries.limedia.fr/ark:/31124/d00c0jkvcxc06ftj/.

[123] À ce sujet, consulter le billet de blog : *Réaliser un mashup de données avec Dataiku DSS et Palladio | Les petites cases* [en ligne]. [s. d.]. [Consulté le 27 février 2020]. Disponible à l'adresse : http://www.lespetitescases.net/realiser-mashup-donnees-Dataiku-DSS-Palladio, ou encore cette vidéo : *Generating Intelligent Multimedia Presentations from Semantic Mashups using OAI-ORE and SMIL* [en ligne]. [s. d.]. [Consulté le 2 février 2020]. Disponible à l'adresse : https://www.youtube.com/watch?v=X6puYLOHQ2c&t=856s.

aujourd'hui une part non négligeable du métier de bibliothécaire. Nous l'avons évoqué plus tôt, les volontés d'interopérabilité se heurtent davantage à un facteur humain qu'à un facteur technologique. Une excellente technologie, qui répond parfaitement à un besoin donné, peut très bien apparaître et tomber dans l'oubli si elle n'est pas reconnue, répandue, si son usage n'est pas généralisé et augmenté.

Afin d'aller dans un des positionnements que nous avons étudiés plus tôt, comme une maîtrise accrue des technologies du web sémantique, ou une transformation des archives ouvertes, il importe que les professionnel-le-s s'emparent véritablement de ces sujets. Parce qu'il est lié à l'informatique, et qu'il implique souvent de passer dans les coulisses, le numérique peut effrayer, mais il est absolument essentiel que les professionnel-le-s de l'information soient partie prenante des évolutions technologiques, des réflexions sur l'évolution des structures. Sans cette implication, le numérique est délégué à des services informatiques qui ont une véritable compétence de développement, mais peu de familiarité – de par leur formation au moins – avec les standards et les pratiques bibliothéconomiques[124].

Pour faire un choix commun, représentatif de la profession, et qui aie donc un véritable pouvoir transformatif, il implique donc d'aller dans le sens d'une démocratisation du numérique dans le monde des bibliothèques. Pour cela, on peut s'appuyer sur des programmes actuellement en cours, notamment celui de la Transition bibliographique que nous avons déjà évoqué : en diffusant le modèle IFLA-LRM, on permet de concevoir différemment les documents et leur description ; en expliquant le lien entre RDA et le web, on sort d'une vision « magique », et on donne un sens à une manière de faire qui peut paraître, de prime abord, une lourdeur supplémentaire sans but précis. À travers des formations de professionnel-le-s à SPARQL, également. Ce protocole permet de rendre, comme on l'a vu, des services à véritable valeur ajoutée au public ; c'est également une manière de résoudre des problèmes en ce qui concerne les personnels pour gérer des collections.

Au-delà d'un programme limité dans le temps, il importe particulièrement de faire du numérique une pierre de touche de l'édifice bibliothéconomique, plutôt qu'une spécialisation. À travers ce mémoire, on a vu des exemples issus du monde du patrimoine, de la lecture publique, du monde de la recherche : chacune de ces communautés a une utilisation de la technologie qui doit être adaptée à ses besoins, et aux projets qui sont les leurs. L'extension du modèle de données que l'on a étudiée au sein de l'université McGill a

[124] À ce sujet, consulter le mémoire de SCHERER, Marc. *Bibliothécaires et informaticiens: convergences ou choc des cultures ?* Villeurbanne, Rhône, France : [s. n.], 2014. École nationale supérieure des sciences de l'information et des bibliothèques.

nécessité, en complément de l'équipe informatique, une bibliothécaire au fait du sujet, des données possédées par la bibliothèque, et des pratiques de recherche des usager-e-s[125]. C'est une expertise extrêmement différente de celle d'un-e développeur-se, mais nécessaire au développement d'une solution satisfaisante.

La participation à des initiatives internationales, comme COAR, IIIF, parmi d'autres, est également une plus-value énorme. En amenant au sein de ces espaces des problématiques particulières, en échangeant avec des collègues et partenaires à un niveau international, on favorise le développement de standards partagés. Dans un moment où la recherche comme le patrimoine sont des espaces de collaboration, fréquemment à l'échelle internationale, il nous semble qu'on ne peut faire l'économie d'aller vers une convergence des pratiques.

Cette convergence ne peut être totale : on a déjà évoqué l'exemple pris par Lars Svensson d'une interopérabilité qui soit suffisante sans être maximaliste. Il importe cependant d'être capable de connaître les données d'une institution et ses besoins.

De ces deux éléments naît la possibilité d'un choix technique, pour répondre à des questions telles que : ce protocole fonctionne-t-il avec mes données ? Répond-il aux besoins de mon institution ? En existe-t-il d'autres qui le font mieux, et si oui, sont-ils répandus dans la communauté ?

De ces deux éléments naît aussi la possibilité de dialoguer et de créer des partenariats avec d'autres acteurs, et aller dans le sens d'une interopérabilité plutôt que d'un outil (ou d'une solution technique). Par cela, nous entendons que l'interopérabilité ne saurait se faire par l'imposition d'une technologie, ou d'un ensemble de technologies supposées répondre à tous les besoins. Par le dialogue et la collaboration, nous construisons l'intéropérabilité nécessaire qu'évoque Lars Svensson en 2014. En ce sens, ce travail ne peut ni ne souhaite prédire une fin du protocole OAI-PMH, ni un nouvel avènement du web sémantique : les deux ont une utilité pour servir les besoins des bibliothèques à l'heure actuelle. Le premier est répandu et sert le besoin de transmettre des métadonnées de manière simple et robuste ; le second peut transformer les services rendus à partir des données que nous possédons, mais doit être répandu au sein de nos communautés professionnelles.

Les professionnel-le-s, dans ce contexte, ont un rôle essentiel de suivi à jouer. Les technologies et standards utilisés en bibliothèque sont, bien souvent, leur apanage quasi-

[125] DESMEULES, Robin Elizabeth. *Entretien sur la création d'un modèle de données compatible avec OAI-PMH au sein de l'université McGill*. 20 février 2020.

exclusif. Les formats MARC, par exemple, n'existent pas hors de notre communauté professionnelle, et même dans ces formats la diversité est grande. OAI-PMH et le Dublin Core ont essaimé en-dehors du monde des bibliothèques, et on retrouve le protocole dans le monde des musées et des archives, dans la communauté des GLAM ; mais on ne peut pas dire qu'au-delà, le protocole soit une des technologies-clés du web. La lente diffusion des technologies dans nos institutions fait qu'au moment où l'on peut dire qu'une technologie est bien implantée dans notre communauté, elle est généralement une technologie de niche, cette niche étant justement la communauté des bibliothèques (au plus, celle des GLAM). Dire que le protocole OAI-PMH est un standard du web est techniquement exact, mais c'est un standard du web de 1999, et pas celui de 2020. Si notre ambition est de rendre les données pertinentes pour les acteurs du web, il convient de ne pas se contenter d'une solution technique, mais de suivre les différents développements en cours, afin de ne pas espérer d'une technologie des effets qu'elle n'est pas, ou plus, capable de produire.

Mais il est évident que le changement est difficile à infuser dans des structures complexes ; sans doute serait-il alors pertinent de chercher à développer une véritable culture de l'expérimentation, en matière de données, dans les différents établissements. Des expérimentations qui peuvent s'opérer en-dehors du système d'information central, celui du catalogue et du SIGB, mais avec les données qu'ils contiennent, et ce dans le plus d'établissements possibles. La BnF et d'autres grandes structures ont des manières de favoriser ce genre d'initiatives, mais il conviendrait d'étendre cela dans des bibliothèques de plus petite taille, universitaires ou de lecture publique. On peut par exemple considérer l'exemple de la BIUS, dont nous avons déjà évoqué la bibliothèque numérique Medica. En 2017-2018, un projet d'ampleur est mené : le versement des portraits numérisés de la BIUS dans Wikimedia Commons[126], la base d'image qui alimente notamment l'encyclopédie collaborative Wikipedia. À l'occasion de ce projet, il y a eu un alignement (partiel) des données et référentiels de la bibliothèque avec le Virtual International Authority File (VIAF), mais aussi Wikidata, des alignements qui ont permis d'améliorer la qualité des métadonnées de la bibliothèque. En termes de plus-value, les images gagnent en visibilité de manière impressionnante : près de deux millions de vues sur Wikimedia pour l'ensemble des images

[126] Voir la communication suivante : BENOIST, David et GHUZEL, Olivier. " Vous voulez mon portrait ? Le voici, dans Commons ! " Le dépôt des portraits numérisés de la Bibliothèque interuniversitaire de Santé dans Wikimédia Commons. Dans : *Journées Wikimédia Culture et Numérique 2019* [en ligne]. Pierrefitte-sur-Seine, France : Wikimédia France, mai 2019. [Consulté le 22 février 2020]. Disponible à l'adresse : https://hal.archives-ouvertes.fr/hal-02144788.

versées en avril 2019, contre moins de dix mille sur la bibliothèque numérique Medica[127]. Enfin, les équipes de la BIUS ont mis en place un type de cercle vertueux avec la communauté : les enrichissements que celle-ci effectue sur les images (identification de personnes, correction d'informations) sont répercutées sur Medica. Cela permet à la bibliothèque de favoriser une information de bonne qualité, souvent correctement sourcée, dans un espace visible et réutilisé, et de profiter des retours de la communauté pour améliorer la qualité de la description de ces contenus. Du point de vue des acteurs du projet, celui-ci a été un succès, et dans les préconisations développées à la fin de la présentation, il est évoqué l'idée d'un référent Wikimedia au sein de la bibliothèque. Cependant, ce projet, une fois achevé, n'a pas donné lieu à une traduction de la tâche de versement et d'alignement dans les fiches de postes des agents.

À notre sens, ce genre de poste pourrait concrétiser des possibilités très intéressantes pour les bibliothèques : les projets de numérisation, en cours dans les bibliothèques depuis parfois vingt ans, continuent, et la question de la visibilité de ces contenus, on l'a vu, continue de se poser. Le fait de les déposer dans des bases qui font partie des communs sur le web apporte une visibilité, et la pratique est également formatrice à la maîtrise des alignements de données et à la question des données liées. Au moment où, à la suite des différentes transitions bibliographiques qui se déroulent dans le monde, l'idée du catalogage dans une instance du logiciel Wikibase[128] se développe, il paraît crucial que les bibliothèques se familiarisent avec l'environnement et ses technologies.

Le lien avec les différents projets de Wikimedia est également intéressant à un autre titre : en observant un mouvement de centralisation des données et des usages, nous avons pointé la difficulté pour les bibliothèques d'être reconnues directement comme des interlocuteurs dans l'environnement numérique. Nous avons étudié Google, qui choisit plutôt Sitemaps qu'OAI-PMH, les différents moteurs de recherche qui utilisent Wikipédia dans leur version du *Knowledge Graph*, et la place croissante que prenait ce dernier. Sans passer par une contractualisation, ou un partenariat avec les acteurs du numérique, il semble peu probable que les résultats des bibliothèques apparaissent jamais dans ce *graph*. Mais des initiatives natives du web comme Wikipédia ou Wikidata sont moissonnées par Google. Une manière d'améliorer la visibilité serait donc de positionner la bibliothèque comme fournisseur de

[127] BENOIST, David et GHUZEL, Olivier, *op.cit.*

[128] À ce sujet, voir les résultats du projet Passage d'OCLC, dans : GODBY, Carol Jean, SMITH-YOSHIMURA, Karen, WASHBURN, Bruce, et al. *Creating library linked data with Wikibase: Lessons learned from project passage*. [S. l.] : OCLC Research, 2019.

données, dans un rôle équivalent au protocole OAI-PMH. C'est un changement de paradigme important, puisqu'il serait question de fournir des données à un organisme qui n'est pas une institution publique, mais le potentiel de contribution à des communs de la connaissance est formidable. Depuis quelques années, la communauté s'empare de la contribution dans ces projets collaboratifs, notamment au cours d'événements ou de chantiers ponctuels (WikiProjects[129], 1Lib1Ref[130]), mais le manque de personnels dédiés fait souvent passer ces projets au stade de chantier ponctuel. Nous avons parlé plus tôt de la volonté de faire sortir les données des bibliothèques des silos que sont les catalogues. De la même manière qu'OAI-PMH était une manière de sortir la donnée d'un silo pour la replacer dans un autre, verser ses données dans une base centralisée comme Wikidata extrait l'information pour la replacer, cette fois-ci, dans un gigantesque silo qui nourrit, au-delà de Wikipedia, une communauté grandissante.

Enfin, favoriser l'expérimentation et l'expertise des données dans les bibliothèques permet, au-delà de choix informés et de contribution à des initiatives internationales, de ne pas prendre de retard sur les nouvelles « vagues » de technologie. Les technologies du web sémantique, ou leurs héritières, ont permis de créer des masses de données structurées, particulièrement utiles pour des opérations d'apprentissage automatique (*machine learning*) et autres technologies de l'intelligence artificielle. Le Metropolitan Museum a par exemple utilisé une intelligence artificielle pour faciliter la création de descriptions d'éléments représentés dans des peintures[131] : ainsi, la description des collections est améliorée, ce qui favorise le nombre de services qu'il est possible de rendre aux publics. Il nous semble que cet exemple représente une possibilité non négligeable pour le monde des bibliothèques : en réseau avec d'autres professions (chercheur-se-s, informaticien-ne-s), s'investir dans le numérique afin de dépasser l'exposition des données des bibliothèques. Si OAI-PMH a été, en son temps, une révélation puisqu'il a permis à des bibliothèques d'exposer leurs données, il nous semble qu'il est nécessaire de s'engager dans une nouvelle « révolution » : celle de

[129] Par exemple ce projet autour des bibliothèques : *Wikipedia:WikiProject Libraries* [en ligne]. [S. l.] : [s. n.], 27 juillet 2019. [Consulté le 22 février 2020]. Disponible à l'adresse : https://en.wikipedia.org/w/index.php?title=Wikipedia:WikiProject_Libraries&oldid=908049333. Page Version ID: 908049333.

[130] *1Lib1Ref* [en ligne]. [S. l.] : [s. n.], 5 février 2020. [Consulté le 22 février 2020]. Disponible à l'adresse : https://fr.wikipedia.org/w/index.php?title=1Lib1Ref&oldid=167097348. Page Version ID: 167097348.

[131] Combining AI and Human Judgment to Build Knowledge about Art on a Global Scale. Dans : *The Metropolitan Museum of Art* [en ligne]. [s. d.]. [Consulté le 22 février 2020]. Disponible à l'adresse : https://www.metmuseum.org/blogs/now-at-the-met/2019/wikipedia-art-and-ai.

l'exploitation de nos données, seul-e-s ou en collaboration, pour inventer les services de demain.

CONCLUSION

QUAND LA POUSSIERE RETOMBE

Quelques vingt ans après la conception théorique du protocole OAI-PMH, quel est le chemin parcouru ? Les bibliothèques ont pris à cœur l'idée de mettre leurs données sur le web. La forme qu'a pris cette idée est, on l'a vu, problématique : *nos* données sont propres à notre communauté professionnelle. Bien souvent, ce que nous mettons sur le web n'est qu'une fraction de l'information dont nous disposons réellement, gêné-e-s en cela par des processus ressentis comme complexes, éloignés peut-être de nos standards métier habituels. Le protocole OAI-PMH, à défaut d'autre chose, a permis de créer une manière simple, sinon simpliste, d'échanger et d'exposer des données entre institutions. Il a été la clé de voûte de nombreux partenariats aux niveaux intra et international. Peut-être aussi a-t-il été un frein dans le développement de solutions alternatives : sa simplicité a garanti son adoption, et son essaimage large ; comparés à ce protocole, les autres projets pourtant développés par la même équipe, n'ont pas eu la même impulsion.

Mais on ne peut pas dire qu'un simple protocole de bibliothèque aie, à lui seul, empêché l'apparition des technologies du web sémantique dans notre communauté professionnelle, de la même manière que leur diffusion actuelle ne sonne pas le glas d'OAI-PMH. Le relativement faible écho des technologies du web sémantique dans l'écosystème du web en général, l'impact extrêmement fort des technologies du web social et de leurs applications ,ont fortement limité l'avènement du web sémantique. Aujourd'hui, certains éléments du web sémantique ont des applications, mais la structure originale de ce que devait être ce web 3.0 est remise en question, notamment par une forte recentralisation. Dans ce cadre, il importe que les solutions techniques que retiennent les bibliothèques correspondent aux objectifs fixés. Le protocole OAI-PMH, par exemple, a sans doute aussi bien fonctionné car, en-dehors de sa robustesse, il était porteur de promesses, en terme de référencement dans les moteurs de recherche et les agrégateurs, qui étaient supposés augmenter logiquement la visibilité des contenus des entrepôts. Le fait d'exposer des données augmente en effet la visibilité, mais l'opacité de la conception des moteurs de recherche, le fait qu'ils développent leurs propres protocole de recherche et d'indexation qui ne se conforment pas avec nos standards, met à mal cette ambition. Le fait que certains contenus ne soient captés que par l'un des agrégateurs

pousse l'usager potentiel à une logique de points d'accès multiples, ce qui est, à l'heure actuelle, contraire aux pratiques de recherche. Ainsi, il convient d'avoir à la fois une vision précise de l'écosystème dans lequel fonctionne la technologie, et une connaissance de la technologie elle-même, de ses apports comme de ses limites. Pour certains projets, OAI-PMH est un outil fonctionnel et utile ; il peut même être amélioré, et être utile au-delà de son cas d'usage originel. Mais, même en étant amélioré, détourné, le protocole reste centré autour des métadonnées, dans un environnement où le point d'attention est désormais la ressource elle-même. Pour des besoins de fouille de texte et de données, d'entraînement de machines, OAI-PMH n'est pas un outil suffisant. Afin d'accommoder ces nouveaux besoins, de nouvelles technologies doivent émerger, ou plutôt se répandre, sachant que les alternatives sont légion. « L'adéquation compte, mais l'utilisation aussi »[132] : parfois, un choix pertinent, et approprié aux besoins doit être mis de côté, au profit d'une solution répandue dans les communautés professionnelles. Cela est vrai, du moins, si dans nos objectifs se pose le problème de la collaboration. Or, l'apparition de la terminologie GLAM, la prise de conscience que les défis, les publics, les problématiques de nos institutions avaient beaucoup en commun, nous impose actuellement de travailler, le plus possible, de manière transversale. Nos ressources connaissent rarement les barrières de nos professions : des bibliothèques conservent des fonds d'archives, des gravures, peintures et objets, et les autres institutions possèdent et conservent, de la même manière, des matériaux que l'on imaginerait à leur place sur les rayonnages d'une bibliothèque. De la même manière, en bibliothèque universitaire, le besoin de travailler avec un grand nombre d'organismes et de personnes diverses impose d'avoir une souplesse et une adaptabilité de nos données, si nous souhaitons qu'elles soient pertinentes.

DES OBJECTIFS AVANT LES SOLUTIONS

En ce sens, il nous semble crucial de ne pas séparer les technologies des objectifs généraux, et de concevoir tout ensemble. Quelles technologies nous permettront d'atteindre les objectifs de la science ouverte, tels que formulés dans les différents plans, propositions, appels ? Afin de trouver ces technologies, nécessaires pour faire communiquer nos contenus, mettre en place des services d'un genre nouveau, il faut que les bibliothèques poursuivent un travail en réseau, qui implique le plus d'établissements possibles. Pour que le changement

[132] DESMEULES, Robin Elizabeth. *Entretien sur la création d'un modèle de données compatible avec OAI-PMH au sein de l'université McGill*. 20 février 2020.

soit véritablement tangible, il importe qu'il soit véritablement commun : le succès d'OAI-PMH tient à son adoption, il convient donc de promouvoir des technologies qui peuvent servir les objectifs nouveaux qui émergent, et qui peuvent être adoptées par l'ensemble de la communauté. De la même manière, dans les communautés patrimoniales, l'exposition de contenus, leur consultation, diffusion et réutilisation est sans cesse en évolution : l'apparition de consortiums internationaux comme IIIF permet l'élaboration de standards partagés par les acteurs du domaine.

Il nous semble que, si la simplicité d'OAI-PMH est également responsable de sa viralité, le monde des bibliothèques ne pourra pas, dans les années à venir, éviter d'embrasser la complexité du paysage, tout du moins dans une certaine mesure. Les problématiques de centralisation ou de décentralisation des contenus sur le web doivent dicter nos pratiques professionnelles, ou du moins conduire à leur adaptation. Pour ce faire, il importe de créer, au sein de nos institutions, des espaces d'expérimentation, de veille et de collaboration avec d'autres acteurs du numérique : organiser une montée en compétence sur ces sujets permet d'envisager les défis à venir comme de véritables opportunités pour nos institutions. À l'heure où la pertinence des bibliothèques est parfois remise en question, puisque « tout est en ligne », il nous semble que l'enjeu de rendre nos contenus pertinents, accessibles, réutilisables, particulièrement de manière numérique, doit être primordial dans la conception de nos missions, et donc de nos établissements.

SOURCES

BERMÈS, Emmanuelle. Entretien sur l'utilisation et l'historique d'OAI-PMH au sein de la Bibliothèque Nationale de France, ainsi que des perspectives dans ce domaine. 27 janvier 2020

COURTIN, Antoine. Entretien sur le protocole OAI-PMH, ses limites et alternatives. 12 juin 2019

DESMEULES, Robin Elizabeth. Entretien sur la création d'un modèle de données compatible avec OAI-PMH au sein de l'université McGill. 20 février 2020

FOUCHER, Tiphaine-Cécile. Entretien sur le fonctionnement de data.bnf.fr, des alignements de données et du futur de la création de données bibliographiques. 21 février 2020

GHUZEL, Olivier. Entretien sur l'utilisation d'OAI-PMH au sein de la Bibliothèque Interuniversitaire de Santé. 31 janvier 2020

MERRIEN, Delphine. Entretien sur Bibliotouch et l'utilisation du protocole OAI-PMH au sein de l'application. 15 mai 2019

POUPEAU, Gautier. Entretien sur les perspectives de l'échange des données et le rôle des professionnel-le-s de l'information à cet égard. 27 février 2020

POUPEAU, Gautier et BERMÈS, Emmanuelle. Entretien croisé sur les problématiques liées à l'exposition de données sur le web et le web sémantique. 10 juin 2019

POUYLLAU, Stéphane. Entretien sur l'utilisation d'OAI-PMH au sein d'Isidore. 20 mai 2019

BIBLIOGRAPHIE

CREER UN SYSTEME DE COMMUNICATION DANS UN CONTEXTE D'ECLOSION D'ARCHIVES OUVERTES

CHAN, Leslie, CUPLINSKAS, Darius, EISEN, Michael, GENOVA, Yana, GUÉDON, Jean-Claude, HAGEMANN, Melissa, HARNAD, Stevan, JOHNSON, Rick, KUPRYTE, Rima, LA MANNA, Manfredi, RÉV, István, SEGBERT, Monika, DE SOUZA, Sidnei, SUBER, Peter et VELTEROP, Jan. *Budapest Open Access Initiative | Read the Budapest Open Access Initiative* [en ligne]. 14 février 2002. [Consulté le 17 janvier 2020]. Disponible à l'adresse : https://www.budapestopenaccessinitiative.org/read

RHEINGOLD, Howard. *The Virtual Community: Homesteading on the Electronic Frontier*. [S. l.] : MIT Press, 23 octobre 2000. ISBN 978-0-262-26110-4. Google-Books-ID: fr8bdUDisqAC

VAN DE SOMPEL, Herbert et LAGOZE, Carl. The Santa Fe Convention of the Open Archives Initiative. *D-Lib Magazine* [en ligne]. Février 2000, Vol. 6, n° 2. [Consulté le 13 janvier 2020]. DOI 10.1045/february2000-vandesompel-oai

WEIBEL, Stuart L. *Metadata: the Foundations of Resource Description* [en ligne]. juillet 1995. [Consulté le 16 janvier 2020]. Disponible à l'adresse : http://www.dlib.org/dlib/July95/07weibel.html

WEIBEL, Stuart L., KUNZE, John A., LAGOZE, Carl et WOLF, Misha. *Request for Comments 2413* [en ligne]. septembre 1998. [Consulté le 15 janvier 2020]. Disponible à l'adresse : https://www.ietf.org/rfc/rfc2413.txt

BibEc main page [en ligne]. 11 décembre 1997. [Consulté le 11 janvier 2020]. Disponible à l'adresse : http://web.archive.org/web/19971211044921/http://netec.mcc.ac.uk/BibEc.html

Economist Jokes [en ligne]. 11 décembre 1997. [Consulté le 11 janvier 2020]. Disponible à l'adresse : http://web.archive.org/web/19971211044513/http://netec.mcc.ac.uk/JokEc.html

Home Page Papers in Economics [en ligne]. 11 décembre 1997. [Consulté le 11 janvier 2020]. Disponible à l'adresse : http://web.archive.org/web/19971211050056/http://netec.mcc.ac.uk/HoPEc.html

NetEc homepage [en ligne]. 11 décembre 1997. [Consulté le 11 janvier 2020]. Disponible à l'adresse : http://web.archive.org/web/19971211044613/http://netec.mcc.ac.uk/

Open Archives Initiative Service Providers [en ligne]. février 2002. [Consulté le 17 janvier 2020]. Disponible à l'adresse : https://web.archive.org/web/20020204155940/http://www.openarchives.org:80/service/listproviders.html

WebEc - WWW Resources in Economics [en ligne]. 11 décembre 1997. [Consulté le 11 janvier 2020]. Disponible à l'adresse : http://web.archive.org/web/19971211045956/http://netec.mcc.ac.uk/WebEc.html

WoPEc main page [en ligne]. 11 décembre 1997. [Consulté le 11 janvier 2020]. Disponible à l'adresse : http://web.archive.org/web/19971211044714/http://netec.mcc.ac.uk/WoPEc.html

UTILISER ET DIFFUSER OAI-PMH

ARMS, Caroline R. Available and useful: OAI at the Library of Congress. *Library Hi Tech* [en ligne]. Juin 2003, Vol. 21, n° 2, p. 129-139. DOI 10.1108/07378830310491899

BERMÈS, Emmanuelle. Tout sur l'OAI. Dans : *Figoblog* [en ligne]. 16 février 2005. [Consulté le 24 juillet 2019]. Disponible à l'adresse : https://figoblog.org/2005/02/16/566/

COCKERILL, Matthew, LAGOZE, Carl et SÉVIGNY, Martin. *[OAI-implementers] Reconsidering mandatory DC in OAI-PMH* [en ligne]. 5 août 2003. [Consulté le 16 janvier 2020]. Disponible à l'adresse : http://www.openarchives.org/pipermail/oai-implementers/2003-August/000953.html

HUNTER, Philip. *OAI and OAI-PMH for absolute beginners: a non-technical introduction [Introduction to OAI and Harvesting]* [en ligne]. 2005. [Consulté le 24 juillet 2019]. Disponible à l'adresse : http://eprints.rclis.org/7143/

LAGOZE, Carl, VAN DE SOMPEL, Herbert, NELSON, Michael et WARNER, Simeon. *Open Archives Initiative - Protocol for Metadata Harvesting - Guidelines for Repository Implementers* [en ligne]. 19 janvier 2005. [Consulté le 31 janvier 2020]. Disponible à l'adresse : https://www.openarchives.org/OAI/2.0/guidelines-repository.htm

OURY, Clément. *Le département de la bibliothèque numérique à la Bibliothèque nationale de France*. 2005, p. 70

POWELL, Andy. *[OAI-implementers] Re: [Dspace-tech] Google Scholar and OAI (fwd)* [en ligne]. 3 février 2005. [Consulté le 24 janvier 2020]. Disponible à l'adresse : http://www.openarchives.org/pipermail/oai-implementers/2005-February/001407.html

VAN DE SOMPEL, Herbert, NELSON, Michael L., LAGOZE, Carl et WARNER, Simeon. Resource Harvesting within the OAI-PMH Framework. *D-Lib Magazine* [en ligne]. Décembre 2004, Vol. 10, nº 12. [Consulté le 10 avril 2019]. DOI 10.1045/december2004-vandesompel

VAN DE SOMPEL, Herbert, YOUNG, Jeffrey A. et HICKEY, Thomas B. Using the OAI-PMH ... Differently. *D-Lib Magazine* [en ligne]. Juillet 2003, Vol. 9, nº 7/8. [Consulté le 10 avril 2019]. DOI 10.1045/july2003-young

WARD, Jewel. Unqualified Dublin Core usage in OAI-PMH data providers. *OCLC Systems & Services: International digital library perspectives* [en ligne]. Mars 2004. [Consulté le 24 juillet 2019]. DOI 10.1108/10650750410527322. World

AIM25 - Online research for archive collections of higher education institutions and learned societies within greater London [en ligne]. [s. d.]. [Consulté le 17 janvier 2020]. Disponible à l'adresse : https://aim25.com/index.stm

California Digital Library [en ligne]. 19 février 2003. [Consulté le 17 janvier 2020]. Disponible à l'adresse : https://web.archive.org/web/20030219022653/http://www2.cdlib.org/

OAIster Home [en ligne]. 17 février 2003. [Consulté le 17 janvier 2020]. Disponible à l'adresse : https://web.archive.org/web/20030217011435/http://www.oaister.org/o/oaister/

Open Archives Initiative Service Providers [en ligne]. 17 avril 2003. [Consulté le 17 janvier 2020]. Disponible à l'adresse : https://web.archive.org/web/20030417164719/http://www.openarchives.org:80/service/listproviders.html

Open Archives Initiative Service Providers [en ligne]. 24 octobre 2005. [Consulté le 17 janvier 2020]. Disponible à l'adresse : https://web.archive.org/web/20051024001606/http://www.openarchives.org:80/service/listproviders.html

LE WEB SEMANTIQUE : PRINCIPES ET PROMESSES EN BIBLIOTHEQUE ET AU-DELA

BERMÈS, Emmanuelle, ISAAC, Antoine et POUPEAU, Gautier. *Le web sémantique en bibliothèque*. Paris, France : Éd. du Cercle de la librairie, 2013. ISBN 978-2-7654-1417-9

BERNERS-LEE, TIM, HENDLER, JAMES et LASSILA, ORA. THE SEMANTIC WEB. *Scientific American*. 2001, Vol. 284, n° 5, p. 34-43. JSTOR

DELESTRE, Nicolas, MALANDAIN, Nicolas et BUSSI, Michel. *Du web des documents au web sémantique*. Bois-Guillaume, France : Éditions KLOG, 2017. ISBN 979-10-92272-18-5

DOERR, Martin, GRADMANN, Stefan, HENNICKE, Steffen, ISAAC, Antoine, MEGHINI, Carlo et SOMPEL, H. The Europeana Data Model (EDM). *World Library and Information Congress: 76th IFLA General Conference and Assembly*. 2010, p. 10-15

HYVÖNEN, Eero. *Publishing and using cultural heritage linked data on the Semantic Web*. San Rafael, Etats-Unis d'Amérique : Morgan & Claypool, cop 2012. ISBN 978-1-60845-998-8

INRIA. MOOC Web sémantique et Web de données. Dans : *FUN-MOOC* [en ligne]. [s. d.]. [Consulté le 24 juillet 2019]. Disponible à l'adresse : //www.fun-mooc.fr/courses/course-v1:inria+41002+self-paced/about

POUPEAU, Gautier. *Les carcans de la pensée hiérarchique et documentaire (2) | Les petites cases* [en ligne]. 7 mai 2009. [Consulté le 2 février 2020]. Disponible à l'adresse : https://www.lespetitescases.net/carcans-de-la-pensee-hierarchique-et-documentaire-2

SAUERMANN, Leo, GMBH, Dfki et CYGANIAK, Richard. *Cool URIs for the semantic web*. Juillet 2011

W3C. *Category:Semantic Web Browser - Semantic Web Standards* [en ligne]. 2009. [Consulté le 6 février 2020]. Disponible à l'adresse : https://www.w3.org/2001/sw/wiki/Category:Semantic_Web_Browser

WELLER, Katrin. *Knowledge representation in the social semantic Web*. Berlin New York : De Gruyter Saur, 2010. Knowledge & information. ISBN 978-3-598-25180-1. TK5105.88815 .W45 2010

Europeana Data Model. Dans : *Europeana Pro* [en ligne]. [s. d.]. [Consulté le 10 février 2020]. Disponible à l'adresse : https://pro.europeana.eu/resources/standardization-tools/edm-documentation

FOAF Vocabulary Specification [en ligne]. [s. d.]. [Consulté le 22 février 2020]. Disponible à l'adresse : http://xmlns.com/foaf/spec/#term_age

SKOS Simple Knowledge Organization System Namespace Document - HTML Variant, 18 August 2009 Recommendation Edition [en ligne]. [s. d.]. [Consulté le 22 février 2020]. Disponible à l'adresse : https://www.w3.org/2009/08/skos-reference/skos.html#mappingRelation

Web sémantique, web de données, liage de données [en ligne]. [s. d.]. [Consulté le 24 janvier 2020]. Disponible à l'adresse : https://www.culture.gouv.fr/Sites-thematiques/Langue-francaise-et-langues-de-France/Politiques-de-la-langue/Langues-et-numerique/Web-semantique-web-de-donnees-liage-de-donnees

SUCCESSION D'OAI-PMH : VERS DE NOUVELLES ARCHIVES INSTITUTIONNELLES ?

ALLINSON, Julie. *Thoughts on Compound Documents* [en ligne]. mai 2005. [Consulté le 2 février 2020]. Disponible à l'adresse : http://www.openarchives.org/ore/documents/CompoundObjects-200705.html

BREEDING, Marshall. *The Impact of OA: Preparing for a New Cycle of Change in Scholarly Publishing* [en ligne]. [s. d.]. [Consulté le 24 juillet 2019]. Disponible à l'adresse : http://www.infotoday.com/cilmag/may19/Breeding--Preparing-for-a-New-Cycle-of-Change-in-Scholarly-Publishing.shtml

ISAAC, Henri. *L'université numérique*. Rapport à Madame Valérie Pécresse, ministre de l'Enseignement Supérieur et de la Recherche. [S. l.] : [s. n.], 2008

KLEIN, Martin et VAN DE SOMPEL, Herbert. Extending Sitemaps for ResourceSync. *arXiv:1305.4890 [cs]* [en ligne]. Mai 2013. [Consulté le 24 juillet 2019]. Disponible à l'adresse : http://arxiv.org/abs/1305.4890. ArXiv: 1305.4890

LUTZ, Jean-François. *Spire : l'archive ouverte de Sciences Po* [en ligne]. 2009. [Consulté le 3 février 2020]. Disponible à l'adresse : https://fr.slideshare.net/jflutz/spire-larchive-ouverte-de-sciences-po

OPEN ARCHIVES INITIATIVE. *ResourceSync Tutorial* [en ligne]. 13:14:15 UTC. [Consulté le 24 juillet 2019]. Disponible à l'adresse : https://www.slideshare.net/OpenArchivesInitiative/resourcesync-tutorial

POYNDER, Richard. Open and Shut?: Q&A with CNI's Clifford Lynch: Time to re-think the institutional repository? Dans : *Open and Shut?* [en ligne]. 22 septembre 2016. [Consulté le 24 juillet 2019]. Disponible à l'adresse : https://poynder.blogspot.com/2016/09/q-with-cnis-clifford-lynch-time-to-re_22.html

SHEARER, Kathleen. *COAR » COAR Updated Feedback on the Guidance on Implementation of Plan S* [en ligne]. [s. d.]. [Consulté le 25 juin 2019]. Disponible à l'adresse : https://www.coar-repositories.org/news-media/coar-feedback-on-the-guidance-on-implementation-of-plan-s/

SHEARER, Kathleen. *COAR » More on the future of repositories – response to Richard Poynder* [en ligne]. [s. d.]. [Consulté le 25 juin 2019]. Disponible à l'adresse : https://www.coar-repositories.org/news-media/more-on-the-future-of-repositories-response-to-richard-poynder/

SVENSSON, Lars. *Jusqu'où l'interopérabilité est-elle nécessaire?* [en ligne]. Montpellier, 20 mai 2014. [Consulté le 20 février 2020]. Disponible à l'adresse : https://pdfslide.net/internet/jusquou-linteroperabilite-est-elle-necessaire.html

TARRANT, David, O'STEEN, Ben, BRODY, Tim, HITCHCOCK, Steve, JEFFERIES, Neil et CARR, Leslie. Using OAI-ORE to Transform Digital Repositories into Interoperable Storage and Services Applications. *The Code4Lib Journal* [en ligne]. Mars 2009, nº 6. [Consulté le 2 février 2020]. Disponible à l'adresse : https://journal.code4lib.org/articles/1062

TAY, Aaron. Rethinking institutional repositories. *Online Searcher*. Mars 2017, Vol. Vol. 41, no 2, nº Vol. 41, no 2, p. 10(6)

VAN DE SOMPEL, Herbert et NELSON, Michael L. Reminiscing About 15 Years of Interoperability Efforts. *D-Lib Magazine* [en ligne]. Novembre 2015, Vol. 21, nº 11/12. [Consulté le 25 janvier 2020]. DOI 10.1045/november2015-vandesompel

VAN DE VELDE, Eric. SciTechSociety: Let IR RIP. Dans : *SciTechSociety* [en ligne]. 24 juillet 2016. [Consulté le 24 juillet 2019]. Disponible à l'adresse : http://scitechsociety.blogspot.com/2016/07/let-ir-rip.html

About COAR. Dans : *COAR* [en ligne]. [s. d.]. [Consulté le 19 février 2020]. Disponible à l'adresse : https://www.coar-repositories.org/about-coar/

COAR Next Generation Repositories: Technologies [en ligne]. [s. d.]. [Consulté le 19 février 2020]. Disponible à l'adresse : https://ngr.coar-repositories.org/technology/

Historique du projet ORI-OAI | ORI-OAI : Valoriser le patrimoine numérique scientifique, pédagogique et documentaire des universités par un réseau de portails communicants [en ligne]. 7 septembre 2011. [Consulté le 6 février 2020]. Disponible à l'adresse : Historique.html

PETALE [en ligne]. 26 décembre 2018. [Consulté le 6 février 2020]. Disponible à l'adresse : https://web.archive.org/web/20181226203412/http://petale.univ-lorraine.fr/index.html

LE WEB SEMANTIQUE EN 2020 : DU BILAN AU PLATEAU DE PRODUCTIVITE

ANDREW W. MELLON FOUNDATION et LIBRARY OF CONGRESS. *LD4P : Linked Data For Production* [en ligne]. 2018. [Consulté le 5 février 2020]. Disponible à l'adresse : https://www.loc.gov/bibframe/news/pdf/ld4p-alamw2018.pdf

BERMÈS, Emmanuelle. L'évolution du modèle d'agrégation de données dans les bibliothèques numériques. Dans : *Figoblog* [en ligne]. 25 mars 2016. [Consulté le 10 avril 2019]. Disponible à l'adresse : https://figoblog.org/2016/03/25/levolution-du-modele-dagregation/

POUPEAU, Gautier. *Les technos du Web sémantique ont-elles tenu leurs promesses ? | Les petites cases* [en ligne]. 6 octobre 2018. [Consulté le 10 avril 2019]. Disponible à l'adresse : http://www.lespetitescases.net/les-technos-du-web-semantique-ont-elles-tenu-leurs-promesses

ROBINEAU, Régis. Comprendre IIIF et l'interopérabilité des bibliothèques numériques. Dans : *Insula* [en ligne]. 8 novembre 2016. [Consulté le 24 juillet 2019]. Disponible à l'adresse : https://insula.univ-lille3.fr/2016/11/comprendre-iiif-interoperabilite-bibliotheques-numeriques/

Google Blog: Webmaster-friendly [en ligne]. 8 juin 2005. [Consulté le 6 février 2020]. Disponible à l'adresse : https://web.archive.org/web/20050608015054/http://googleblog.blogspot.com/2005/06/webmaster-friendly.html

Home — IIIF | International Image Interoperability Framework [en ligne]. [s. d.]. [Consulté le 28 juin 2019]. Disponible à l'adresse : https://iiif.io/

Wikidata:WikiProject Freebase - Wikidata. Dans : *wikidata.org* [en ligne]. 17 décembre 2014. [Consulté le 11 février 2020]. Disponible à l'adresse : https://www.wikidata.org/wiki/Wikidata:WikiProject_Freebase

PANORAMA DES BIBLIOTHEQUES EN 2020 : FOURNISSEUSES DE DONNEES, DE SERVICES ET DE RESSOURCES

BIBLIOTHÈQUE NATIONALE DE FRANCE, Direction des Services et des Réseaux, Département de la Coopération. *Les partenaires de Gallica en 2017* [en ligne]. 2017. [Consulté le 14 février 2020]. Disponible à l'adresse : https://www.bnf.fr/sites/default/files/2018-11/partenaires_gallica.pdf

BOUCHARD, Ariane. Présentation du projet CORPUS à la BnF. Dans : *Web Corpora* [en ligne]. [s. d.]. [Consulté le 18 février 2020]. Disponible à l'adresse : https://webcorpora.hypotheses.org/119

ENSSIB. *OAI-PMH | Dictionnaire de l'Enssib* [en ligne]. [s. d.]. [Consulté le 28 juin 2019]. Disponible à l'adresse : https://www.enssib.fr/le-dictionnaire/oai-pmh

INTERNATIONAL ORGANIZATION FOR STANDARDIZATION. ISO 19115-1:2014. Dans : *ISO* [en ligne]. [s. d.]. [Consulté le 24 janvier 2020]. Disponible à l'adresse : http://www.iso.org/cms/render/live/en/sites/isoorg/contents/data/standard/05/37/53798.html

LECLAIRE, Céline. *La BnF à l'âge de la multitude* [en ligne]. 1 janvier 2017. [Consulté le 10 mai 2019]. Disponible à l'adresse : http://bbf.enssib.fr/consulter/bbf-2017-13-0142-001

PARIENTÉ, Jonathan et FERRER, Maxime. 26 000 votants, 70 000 livres, 5 choix : comment le palmarès des 101 romans des lecteurs est né. *Le Monde.fr* [en ligne]. 27 décembre 2019. [Consulté le 18 février 2020]. Disponible à l'adresse : https://www.lemonde.fr/culture/article/2019/12/27/26-000-votants-70-000-livres-5-choix-comment-le-palmares-des-101-romans-des-lecteurs-est-ne_6024215_3246.html

ROCHE, Mélanie. *En attendant « le jour […] où il n'y aura plus de catalogue à faire » : une histoire matérielle des catalogues de bibliothèque (1789 – 1993)* [en ligne]. Mémoire d'étude. [S. l.] : ENSSIB, janvier 2014. [Consulté le 24 janvier 2020]. Disponible à l'adresse : https://www.enssib.fr/bibliotheque-numerique/documents/64118-en-attendant-le-jour-ou-il-n-y-aura-plus-de-catalogue-a-faire-une-histoire-materielle-des-catalogues-de-bibliotheque-1789-1993.pdf

ROCHE, Mélanie et PEYRARD, Sébastien. *À défaut d'enterrement : les défis et les promesses de l'INTERMARC nouvelle génération*. 2018

Archivage pérenne | CINES [en ligne]. [s. d.]. [Consulté le 6 février 2020]. Disponible à l'adresse : https://www.cines.fr/archivage-perenne/

Bibliothèque numérique Medica — BIU Santé, Paris [en ligne]. [s. d.]. [Consulté le 10 février 2020]. Disponible à l'adresse : https://www.biusante.parisdescartes.fr/histoire/medica/index.php

Bibliotouch [en ligne]. [s. d.]. [Consulté le 20 février 2020]. Disponible à l'adresse : https://bibliotouch.enssib.fr/#/

Calames : nouveau service OAI-PMH. Dans : *FIL'ABES* [en ligne]. 10 octobre 2017. [Consulté le 31 janvier 2020]. Disponible à l'adresse : https://fil.abes.fr/2017/10/10/calames-nouveau-service-oai-pmh/

clélie histoire romaine - Recherche Google. Dans : *Google* [en ligne]. 18 février 2020. [Consulté le 18 février 2020]. Disponible à l'adresse : https://www.google.com/search?sxsrf=ACYBGNTw0BsgfjisX4iwnJCiY_WHRYvTcQ%3A1582020839895&ei=57hLXtKZNp6SjLsPy46n0Ag&q=cl%C3%A9lie+histoire+romaine&oq=cl%C3%A9lie&gs_l=psy-ab.3.3.0l3j0i67j0l6.20939.21968..23945...0.2..0.322.1145.0j4j1j1.....0....1..gws-wiz.....0i71j35i39j0i131.HuwlInMdCdE

Historique des travaux français sur RDA. Dans : *www.transition-bibliographique.fr* [en ligne]. [s. d.]. [Consulté le 18 février 2020]. Disponible à l'adresse : https://www.transition-bibliographique.fr/enjeux/historique-travaux-francais-rda/

« Médecine arabe »— Medica — BIU Santé, Paris [en ligne]. [s. d.]. [Consulté le 14 février 2020]. Disponible à l'adresse : https://www.biusante.parisdescartes.fr/histoire/medica/resultats/?intro=arabe&statut=charge

Présentation du modèle de données. Dans : *data.bnf.fr* [en ligne]. [s. d.]. [Consulté le 10 février 2020]. Disponible à l'adresse : https://data.bnf.fr/opendata#Ancre5

Table des planches du « Recueil de belles maisons, hôtels, châteaux exécutés en Lorraine ». Dans : *galeries.limedia.fr* [en ligne]. [s. d.]. [Consulté le 19 février 2020]. Disponible à l'adresse : https://galeries.limedia.fr/ark:/31124/d00c0jkvcxc06ftj/

Transférer le dépôt vers arXiv – HAL Documentation [en ligne]. [s. d.]. [Consulté le 6 février 2020]. Disponible à l'adresse : https://doc.archives-ouvertes.fr/deposer/transfert-hal-arxiv/

[Trône de Dagobert]. Dans : *Gallica* [en ligne]. 0900 750. [Consulté le 31 janvier 2020]. Disponible à l'adresse : https://gallica.bnf.fr/ark:/12148/btv1b55010079x

STANDARDS, NORMES ET PROTOCOLES DIVERS

IFLA -- Functional Requirements for Bibliographic Records [en ligne]. [s. d.]. [Consulté le 22 février 2020]. Disponible à l'adresse : https://www.ifla.org/publications/functional-requirements-for-bibliographic-records

IFLA -- IFLA Library Reference Model (LRM) [en ligne]. [s. d.]. [Consulté le 20 février 2020]. Disponible à l'adresse : https://www.ifla.org/publications/node/11412

Le protocole Z39.50. Dans : *BnF - Site institutionnel* [en ligne]. [s. d.]. [Consulté le 27 janvier 2020]. Disponible à l'adresse : https://www.bnf.fr/fr/le-protocole-z3950

LOMFR (Learning Object Metadata). Dans : *éduscol, le site des professionnels de l'éducation* [en ligne]. [s. d.]. [Consulté le 14 février 2020]. Disponible à l'adresse : https://eduscol.education.fr/numerique/dossier/archives/metadata/ressources-educatives-numeriques/lomfr-learning-object-metadata

SPAR (Système de Préservation et d'Archivage Réparti). Dans : *BnF - Site institutionnel* [en ligne]. [s. d.]. [Consulté le 14 février 2020]. Disponible à l'adresse : https://www.bnf.fr/fr/spar-systeme-de-preservation-et-darchivage-reparti

TEF : Métadonnées des thèses françaises [en ligne]. [s. d.]. [Consulté le 14 février 2020]. Disponible à l'adresse : http://www.abes.fr/abes/documents/tef/

ANNEXES

GRILLE UTILISEE POUR LES ENTRETIENS

Grille d'entretien

L'institution et les buts recherchés

Présentation rapide de l'institution – quels sont les buts en terme d'exposition de contenus sur le web ?

La mise en place du protocole

- *Quel a été votre premier contact avec le protocole OAI-PMH (si pertinent) ?*

Le fonctionnement

- *Qui moissonne les données que vous exposez ?*
- *Quel est le schéma des données exposées ?*
- *Comment mesurez-vous l'utilisation de votre moissonnage ? Si oui, combien d'institutions ? Est-ce un indicateur utilisé (rapport de fin d'années, etc.…) ?*
- *Moissonnez-vous des données ? Si oui, de quels fournisseurs ?*
- *Est-ce que ce qu'il convient de faire la médiation de l'existence de ces données/métadonnées ? Si oui, auprès d'un réseau déjà existant ? Lequel ?*

Les limites et alternatives éventuelles

Les limites

- *Quelles sont, de votre expérience, les limites du protocole OAI-PMH ?*

Les alternatives éventuelles

- *Quelles sont, de votre expérience, les alternatives qu'il serait possible de mettre en place ?*
- *Des réflexions ou expérimentations ont-elles été entreprises en ce sens ?*

Le web sémantique

- *Quelle est votre familiarité avec les technologies du web sémantique (si pertinent) ?*
- *Est-on à « l'heure du web sémantique » ?*
- *De quelle manière votre institution a-t-elle expérimenté, ou expérimente encore, avec le web sémantique ?*

Discussion libre

Quel est votre point de vue sur le web à l'heure actuelle, et ses possibles évolutions ? Quelle est la place des bibliothèques dans ce paysage ? Quels points de vigilance doit-on garder à l'esprit ?

GLOSSAIRE

Archive ouverte : en anglais, *repository* (entrepôt). Réservoir qui héberge des données liées, historiquement, à la recherche et l'enseignement (souvent des articles scientifiques, thèses, communication dans des congrès...), et dont l'accès se veut le plus ouvert possible, sans barrière de coût. Le concept apparaît dans la fin des années 1980 pour améliorer et accélérer la communication scientifique, dans un contexte de l'apparition de la notion de l'accès ouvert (*open access*). Depuis, les archives ouvertes se sont diversifiées : elles peuvent être nationales, disciplinaires ou institutionnelles.

Archive institutionnelle : réservoir lié à une institution (université, école, structure de recherche...). Historiquement, les archives institutionnelles se sont multipliées des années 1990 aux années 2010. Mais aujourd'hui, en France, des questions de coûts trop élevés pour une institution, ainsi que des considérations stratégiques sur la science ouverte font reculer les archives institutionnelles au profit d'une solution nationale. Dans d'autres zones du monde, particulièrement l'Amérique du Nord, les archives institutionnelles sont remises en question en ce qui concerne leur efficacité : problèmes d'attractivité, de coût, manque d'usage par les chercheur-se-s sont fréquemment cités.

Accès ouvert : en anglais, *open access*. Formalisé en 2002 par la déclaration de Budapest, le concept de l'accès ouvert favorise l'accès libre à la production scientifique, en passant par deux « voies ». La « voie verte » est celle de l'auto-archivage : une-e chercheur-se dépose son article dans une archive ouverte, et celui-ci est ainsi librement accessible pour la communauté. La « voie dorée » inaugure un système où l'auteur-rice paie pour être publié dans une revue, qui met ensuite l'article à disposition de la communauté, sans abonnement.

Web sémantique : ensemble de standards, de formats et de technologies qui forment le « web 3.0 », supposé remplacer le web dit « social ». La logique essentielle du web sémantique est de passer d'un web de documents à un web de ressources. Dans cette logique, les machines sont capables de comprendre les requêtes formulées dans un langage naturel,

et de fournir les ressources recherchées par les usager-e-s. Mais le succès du web social, la complexité des technologies du web sémantique ont remis en question l'avènement de ce dernier.

Web social : ensemble de standards, de formats et de technologies qui forment le « web 2.0 ». Caractérisé par un engagement plus fort des usager-e-s du web, le web social pousse à une position de contribution. L'apparition des réseaux sociaux, des wikis, et l'engouement qu'ils suscitent ont porté les acteurs commerciaux du web à investir massivement dans ce modèle.

Métadonnée : la métadonnée est, littéralement, une données sur une donnée. En les multipliant, on peut décrire des contenus. La notice d'un livre, dans une bibliothèque, est constituée de métadonnées (le nom de l'auteur, la langue de l'ouvrage, la date de publication...). L'informatisation des contenus étend le rôle des métadonnées : elles ne sont plus simplement descriptives, mais aussi techniques, administratives et structurelles.

OAI-PMH : protocole informatique permettant d'échanger des métadonnées. Le protocole fonctionne avec un fournisseur de données (ou plusieurs), et un fournisseur de services (ou plusieurs). Le fournisseur de données expose, dans un entrepôt informatique, les métadonnées correspondant à des ressources ; le fournisseur de services moissonne ces métadonnées, et peut donc les exposer. Le protocole est particulièrement utile dans le cadre, par exemple, d'une archive ouverte nationale : en moissonnant plusieurs entrepôts (correspondant par exemple à des universités), une archive ouverte nationale peut intégrer les ressources de ces entrepôts dans son catalogue.

Centralisation/décentralisation sur le web : le web est, au moment de sa création, une structure décentralisée. Il n'y a pas de registre centralisé des liens entre les documents. On peut créer, depuis une page, un lien vers une autre page sans avoir besoin d'établir une entente. Cette conception a permis une croissance du web dans les premières années de sa création. Mais si la structure est décentralisée, les usages se recentralisent énormément depuis quelques années, avec de grandes plateformes qui concentrent l'essentiel du trafic. Les questions de coût liées au maintien d'un réseau décentralisé favorisent également une recentralisation des données.

De la métadonnée à la ressource : pour répondre aux problématiques actuelles (fouille de texte, *machine learning*), l'exposition des métadonnées ne suffit pas. Il importe d'avoir accès aussi et surtout aux contenus, afin de les exploiter. Ce « tropisme de la ressource » demande une évolution des technologies et des usages afin d'accommoder les nouveaux besoins.

Patrimoine numérique : ensemble des contenus patrimoniaux exposés sur le web, numérisés ou nativement numériques. Le but des établissements qui gèrent le patrimoine numérique est de permettre l'accès du public à ces contenus. Ce public est, en principe, plus large que celui de l'enseignement supérieur et de la recherche, puisque le patrimoine numérique peut aussi bien être un matériau de recherche qu'un ensemble de contenus à destination du grand public (par exemple, la presse locale ancienne). Il convient donc de donner accès aux documents de manière simple, mais aussi de pouvoir accommoder des demandes plus spécialisées.

TABLE DES MATIERES