

Diplôme de conservateur de bibliothèque

Mémoire de recherche / mars 2021

**Métadonnées pour la science ouverte :  
rôle et action des bibliothèques  
et des professionnels de l'information  
scientifique et technique**

**Vincent Richard**

Sous la direction de David Aymonin  
Directeur de l'Agence bibliographique de l'enseignement supérieur (Abes)



## ***Remerciements***

*Je remercie d'abord David Aymonin pour sa disponibilité et ses conseils avisés tout au long de la préparation de ce mémoire.*

*Je remercie ensuite les professionnels avec qui je me suis entretenu et qui m'ont fait découvrir leurs méthodes de travail, les outils qu'ils utilisent ou les projets qu'ils pilotent : Nicolas Alarcon (SCD de l'université de La Réunion), Luc Bellier et Cédric Mercier (SCD de l'université Paris-Saclay), Éric Jeangirard et Emmanuel Weisenburger (ministère de l'Enseignement supérieur, de la Recherche et de l'Innovation), Frédérique Joannic-Seta (département des Métadonnées de la BnF), Yann Nicolas (Abes) et Thomas Parisot (Cairn).*

*Je remercie également Lou Delaveau et Louis Tisserand, de la promotion DCB 29, pour les retours d'expérience qu'ils ont pris le temps de me présenter.*

**Résumé :** À l'heure de la science ouverte et du web sémantique, les professionnels de l'information scientifique et technique font face à de nombreux défis dans la gestion des métadonnées des productions académiques, qui constituent le socle de la recherche d'information scientifique et de l'accès aux publications au niveau mondial : traitement des données des éditeurs, archives ouvertes, édition en libre accès, gestion et politique des registres d'identifiants, bases de données de publications, devenir des bases de citations, enjeu de la création d'un index mondial ouvert des publications, avancée en matière de graphes de données liées... Nous en présentons les enjeux stratégiques, institutionnels et techniques. Nous tentons également d'identifier les évolutions en cours et souhaitables en termes de compétences, d'outils technologiques et de gouvernance, et de définir le rôle que peuvent jouer les bibliothèques pour renforcer la qualité et l'ouverture des métadonnées.

**Descripteurs** (concepts Rameau) : Métadonnées – Bibliothèques, Web sémantique, Digital Object Identifiers, Normes, Bases de données, Édition en libre accès, Bibliothèques universitaires, Information scientifique, Sciences – Documentation, Langages documentaires

**Abstract :** In the era of open science and semantic web, information science professionals face many challenges in the management of research output metadata, which are the cornerstone of scientific information search and of access to scientific publications at a global level: publishers' data wrangling, open archives, open access publishing, strategy and management of identifier registries, academic databases, future of citation databases, challenge of creating a global open index of publications, linked data graph breakthrough... We present the strategical, institutional and technical issues of these challenges, we try to identify current and desirable evolutions in terms of skills, technological tools and governance, and to define the role that libraries may play to reinforce the quality and openness of metadata.

**Keywords** (LCSH) : Library metadata, Semantic web, Digital Object Identifiers, Standards, Databases, Open access publishing, Academic libraries, Communication in science, Scientific literature, Subject headings



Cette création est mise à disposition selon le Contrat :  
« **Paternité-Pas d'Utilisation Commerciale-Pas de Modification 4.0 France** »  
disponible en ligne <http://creativecommons.org/licenses/by-nc-nd/4.0/deed.fr> ou par  
courrier postal à Creative Commons, 171 Second Street, Suite 300, San Francisco,  
California 94105, USA.

## Sommaire

<b>SIGLES ET ABRÉVIATIONS .....</b>	<b>8</b>
<b>INTRODUCTION.....</b>	<b>13</b>
<b>PARTIE A : LES MÉTADONNÉES DES RESSOURCES SCIENTIFIQUES ET LEUR ÉCOSYSTÈME .....</b>	<b>16</b>
<b>Chapitre 1 : un écosystème complexe.....</b>	<b>16</b>
<i>Des métadonnées diverses.....</i>	<i>16</i>
<i>Des acteurs et enjeux multiples dans un flux de métadonnées.....</i>	<i>17</i>
<i>Focus : les métadonnées de ressources électroniques .....</i>	<i>20</i>
<i>Les principes Fair : quatre piliers pour une bonne gestion des métadonnées .....</i>	<i>21</i>
<i>Un arsenal d'outils pour décrire les ressources scientifiques de façon standard.....</i>	<i>23</i>
<i>Bilan d'étape .....</i>	<i>27</i>
<b>Chapitre 2 : gouvernance de la science ouverte et des métadonnées de la recherche.....</b>	<b>28</b>
<i>En France : un cadre incitatif à l'ouverture des données et un écosystème d'acteurs.....</i>	<i>28</i>
<i>En Europe, une série d'initiatives articulées pour la science ouverte ..</i>	<i>33</i>
<i>Dans le monde, un réseau d'institutions œuvrant à la structuration et à l'ouverture des métadonnées .....</i>	<i>34</i>
<i>Un écosystème international dynamique, dans lequel la France est encore peu impliquée.....</i>	<i>40</i>
<i>Vers une base de données ouverte et mondiale des publications scientifiques ? .....</i>	<i>43</i>
<i>Bilan d'étape .....</i>	<i>49</i>
<b>PARTIE B : LA GESTION DES MÉTADONNÉES DANS LES ÉTABLISSEMENTS DE L'ESR .....</b>	<b>51</b>
<b>Chapitre 3 : l'enjeu des identifiants et référentiels.....</b>	<b>51</b>
<i>Identifiants pérennes et interopérabilité sur le web sémantique .....</i>	<i>51</i>
<i>Quels critères d'identifiants de qualité ? .....</i>	<i>53</i>
<i>Identifiants et référentiels .....</i>	<i>54</i>
<i>Quelles actions sont envisagées en France pour développer l'usage des identifiants et référentiels ? .....</i>	<i>55</i>
<i>Bilan d'étape .....</i>	<i>61</i>
<b>Chapitre 4 : métadonnées et « inside-out collection ».....</b>	<b>62</b>
<i>Quelle gestion des métadonnées en archives ouvertes ? .....</i>	<i>62</i>
<i>Quelles métadonnées pour les publications scientifiques en accès ouvert ?.....</i>	<i>68</i>

<i>Bilan d'étape</i> .....	71
<b>Chapitre 5 : les traitements automatisés de données au service de la qualité</b> .....	<b>72</b>
<i>La curation des métadonnées des éditeurs</i> .....	72
<i>Le travail sur les données de l'ESR français et des établissements de l'ESR</i> .....	76
<i>Bilan</i> .....	88
<b>CONCLUSION</b> .....	<b>89</b>
<b>ANNEXE</b> .....	<b>93</b>
<b>BIBLIOGRAPHIE</b> .....	<b>108</b>
<b>TABLE DES ILLUSTRATIONS</b> .....	<b>121</b>
<b>TABLE DES MATIÈRES</b> .....	<b>122</b>

## *Sigles et abréviations*

- Abes : Agence bibliographique de l'enseignement supérieur
- ABF : Association des bibliothécaires de France
- ADBU : Association française des directeurs et personnels de direction des bibliothèques universitaires et de la documentation
- AMI : appel à manifestation d'intérêt
- ANR : Agence nationale de la recherche
- Ansi : American National Standards Institute
- API : Application Programming Interface
- APC : Article Processing Charges (frais de traitement des articles)
- ARK : Archival Resource Key
- Auréal : Accès unifié aux référentiels HAL
- Bacon : Base de connaissance nationale
- BASE : Bielefeld Academic Search Engine
- Bibframe : Bibliographic Framework Initiative
- Bibo : Bibliographic Ontology
- BIS : Bibliothèque interuniversitaire de la Sorbonne
- BnF : Bibliothèque nationale de France
- BSO : Baromètre de la science ouverte
- Bulac : Bibliothèque universitaire des langues et civilisations
- BU : bibliothèque universitaire
- Calames : Catalogue en ligne des archives et des manuscrits de l'enseignement supérieur
- CAPSH : Comité pour l'accessibilité aux publications en sciences et humanités
- Casrai : Consortia Advancing Standards in Research Administration Information
- CBU : contrôle bibliographique universel
- CC : Creative Commons (licences)
- CCFR : Catalogue collectif de France
- CCSD : Centre pour la communication scientifique directe
- CHU : centre hospitalier universitaire
- Cieps : Centre international d'enregistrement des publications en série
- Cito : Citation Typing Ontology
- CMS : Content Management System (système de gestion de contenu)
- CNRS : Centre national de la recherche scientifique
- Coci : OpenCitations Index of Crossref open DOI-to-DOI citations
- Comue : communautés d'universités et établissements



Core : Connecting Repositories  
CoSO : Comité pour la science ouverte  
CRedit : Contributor Roles Taxonomy  
CRIS : Current Research Information Systems  
CrosCI : Crowdsourced Open Citations Index  
DCMI : Dublin Core Metadata Terms  
DOAJ : Directory of Open Access Journals  
Dorandum : Données de la recherche apprentissage numérique  
DOI : Digital Object Identifier  
EAD : Encoded Archival Description  
EHESS : École des hautes études en sciences sociales  
eISSN : Electronic International Standard Serial Number  
EOSC : European Open Science Cloud  
EPST : Établissement public à caractère scientifique et technologique  
ESR : enseignement supérieur et recherche  
Fair : Facile à trouver, Accessible, Interopérable et Réutilisable (Findable, Accessible, Interoperable, Reusable)  
fMeSH : version française de Medical Subject Headings  
FNSO : Fonds national pour la science ouverte  
Foaf : Friend of a friend (ontologie)  
Frantiq : Fédération et ressources sur l'Antiquité  
FRBR : Functional Requirements for Bibliographic Records (Fonctionnalités requises des notices bibliographiques)  
GOKb : Global Open Knowledgebase  
Grid : Global Research Identifier Database (base de données mondiale des identifiants de la recherche)  
Grobid : Generation of Bibliographic Data  
H2020 : Horizon 2020 (programme européen pour la recherche et le développement pour la période 2014-2020)  
HAL : Hyper articles en ligne (archive ouverte pluridisciplinaire nationale française)  
I4OA : Initiative for Open Abstracts  
I4OC : Initiative for Open Citations  
IdHAL : identifiant auteur de l'archive ouverte HAL  
IdRef : Identifiants et référentiels pour l'enseignement supérieur et la recherche  
Ifla : International Federation of Library Associations and Institutions (Fédération internationale des associations et institutions de bibliothèques)  
Ifla-LRM : IFLA Library Reference Model (modèle de référence IFLA pour les bibliothèques)  
INA : Institut national de l'audiovisuel

Inist : Institut de l'information scientifique et technique  
Inrae : Institut national de recherche pour l'agriculture, l'alimentation et l'environnement  
Insee : Institut national de la statistique et des études économiques  
ISBD : International Standard Bibliographic Description  
ISBN : International Standard Book Number  
Isni : International Standard Name Identifier  
ISO : International Organization for Standardization  
ISSN : International Standard Serial Number  
IST : information scientifique et technique  
Jisc : originellement Joint Information Systems Committee  
Json : JavaScript Object Notation  
Json-LD : JavaScript Object Notation for Linked Data  
Jats : Journal Article Tag Suite (standard de description bibliographique)  
KBart : Knowledge bases and related tools  
LCSH : Library of Congress Subject Headings  
Liber : Ligue des bibliothèques européennes de recherche  
LilloA : Lille Open Access  
Lirmm : Laboratoire d'informatique, de robotique et de microélectronique de Montpellier  
LOD : Linked Open Data (données ouvertes liées)  
MAA : manuscrit auteur accepté  
Mads : Metadata Authority Description Schema  
Marc : Machine-Readable Cataloging  
MeSH : Medical Subject Headings  
Mesri : ministère de l'Enseignement supérieur, de la Recherche et de l'Innovation  
Mets : Metadata Encoding and Transmission Standard  
MNHN : Muséum national d'histoire naturelle  
Mods : Metadata Object Description Schema  
NIH : National Institutes of Health (États-Unis)  
Niso : National Information Standards Organization (États-Unis)  
NLM : National Library of Medicine (États-Unis)  
NLP : Natural Language Processing (traitement automatique des langues)  
OA : Open Access  
OS : Open Science  
OAI-PMH : Open Archives Initiative Protocol for Metadata Harvesting  
Oatao : Open Archive Toulouse Archive Ouverte  
OCI : OpenCitations Identifier

OCLC : Online Computer Library Center (originellement Ohio College Library Center)

ODC-BY : Open Data Commons Attribution License

Onix : Online Information eXchange

Onix-PL : ONIX for Publications Licenses

OpenDOAR : Directory of Open Access Repositories

Operas : Open Scholarly Communication in the European Research Area for Social Sciences and Humanities

Orcid : Open Researcher and Contributor ID (identifiant ouvert pour chercheur et contributeur)

Ortolang : Outils et ressources pour un traitement optimisé de la langue

OWL : Web Ontology Language

Pactols : Peuples et cultures, anthroponymes, chronologie, toponymes, œuvres, lieux et sujets (thésaurus)

Panist : Plateforme d'archivage national de l'information scientifique et technique

PNSO : Plan national pour la science ouverte

Premis : Preservation Metadata Implementation Strategies

Rameau : Répertoire d'autorité matière encyclopédique et alphabétique unifié

RDA : Ressources : description et accès (Resource Description and Access)

RDF : Resource Description Framework

RDFa : Resource Description Framework in Attributes

RDFS : Resource Description Framework Schema

Référens III : Référentiel des emplois-types de la recherche et de l'enseignement Supérieur III

Reiso : Réseau d'experts internationaux de la science ouverte

Repec : Research Papers in Economics

RNSR : Répertoire national des structures de recherche

ROR : Research Organization Registry (registre des organisations de recherche)

Sampra : Software for Analysis and Management of Publications & Research Assessment

Scoss : Global Sustainability Coalition for Open Science Services

SHS : sciences humaines et sociales

SIGB : système intégré de gestion de bibliothèque

Sirene : Système informatique pour le répertoire des entreprises et des établissements

Siret : Système informatique pour le répertoire des entreprises sur le territoire

Skos : Simple Knowledge Organization System

Sparc Europe : Scholarly Publishing and Academic Resources Coalition (Coalition de l'édition savante et des ressources académiques)

Sparql : SPARQL Protocol and RDF Query Language

SRU : Search/Retrieve via URL  
STM : Science, technologies et médecine (ou Scientifique, technique et médical)  
Sudoc : Système universitaire de documentation  
Sudoc-PS : Système universitaire de documentation-Publications en série  
Sword : Simple Web-service Offering Repository Deposit  
TEI : Text Encoding Initiative  
TGIR : très grande infrastructure de recherche  
Unesco : United Nations Educational, Scientific and Cultural Organization  
URI : Uniform Resource Identifier  
UTF-8 : Universal Character Set Transformation Format - 8 bits  
UVSQ : université de Versailles-Saint-Quentin-en-Yvelines  
Viaf : Virtual International Authority File  
W3C : World Wide Web Consortium  
WoS : Web of Science  
XML : Extensible Markup Language  
XSLT : eXtensible Stylesheet Language Transformations

## INTRODUCTION

---

Les métadonnées sont des données sur des données, des informations sur un contenu. Dans le domaine des bibliothèques, elles sont des éléments décrivant des objets documentaires, dénommées historiquement « information secondaire » sur un document primaire<sup>1</sup>. En cela, elles sont des outils essentiels dans l'économie des ressources scientifiques. Grâce à elles, les publications papier ou numériques, textuelles ou audiovisuelles, cartes, archives ou encore jeux de données sont rendus identifiables dans les catalogues et bases de données, et sur le web, et trouvent davantage d'utilisateurs pour les consulter, les réutiliser et les citer.

Si les métadonnées sont au cœur du monde des bibliothèques depuis au moins Callimaque de Cyrène<sup>2</sup>, le développement depuis deux décennies du web sémantique redéfinit les enjeux qui les sous-tendent. Il ne s'agit plus seulement de rendre les ressources d'une bibliothèque découvrables dans un catalogue, mais aussi d'exposer les données (d'autorité et bibliographiques) sur le web, dans l'écosystème de données ouvertes liées (*linked open data*, LOD) qui constituent le fondement du web sémantique. Cela implique de passer d'une logique de documents à une logique d'entités, pour fournir d'abord de l'information sur des personnes, des œuvres, des concepts, etc., et non simplement sur des documents. Le monde des bibliothèques est acteur de ce mouvement notamment par la Transition bibliographique<sup>3</sup>, qui s'appuie sur le modèle conceptuel IFLA-LRM, fondé notamment sur les entités Œuvre, Expression, Manifestation et Items, ainsi que Agent, Res, Nomen, etc.<sup>4</sup> Les bibliothèques passent ainsi de la standardisation de notices pour les échanger à la structuration des données pour les partager<sup>5</sup>.

La question des métadonnées en science ouverte participe de cette évolution, mais des problématiques spécifiques s'y ajoutent. Le questionnement est double : 1) comment attacher aux productions en accès ouvert des métadonnées de qualité ? Mais aussi : 2) comment faire en sorte d'ouvrir les métadonnées des productions scientifiques (y compris celles des productions en accès fermé) ? Car si l'idéal de la science ouverte est d'avoir des publications (ainsi que des données de la recherche et des logiciels) ouvertes, avec à la fois le plein-texte et les métadonnées librement accessibles, la mise à disposition au minimum de métadonnées de qualité, en accès ouvert et interopérables, est un impératif pour en permettre l'exploration, la découverte, la citation, l'accès et l'archivage.

Ces objectifs dépendent notamment des pourvoyeurs de métadonnées. On peut distinguer deux sources majeures des publications scientifiques et donc des métadonnées associées : les éditeurs d'une part, et les institutions de recherche d'autre part. Et au sein de ces institutions, les chercheurs eux-mêmes, les laboratoires, mais aussi les bibliothèques.

---

<sup>1</sup> « L'information secondaire du document primaire : Format Marc ou SGML », Catherine Lupovici, : Bulletin d'informations de l'ABF, n° 174, 1997, <https://www.enssib.fr/bibliotheque-numerique/documents/45355-1-information-secondaire-du-document-primaire.pdf>. Consulté le 16 février 2021

<sup>2</sup> « Catalogue de la bibliothèque d'Alexandrie », Wikipédia, [https://fr.wikipedia.org/wiki/Catalogue\\_de\\_la\\_biblioth%C3%A8que\\_d%27Alexandrie](https://fr.wikipedia.org/wiki/Catalogue_de_la_biblioth%C3%A8que_d%27Alexandrie). Consulté le 15 février 2021

<sup>3</sup> « À la Une de la Transition bibliographique », Transition bibliographique, <https://www.transition-bibliographique.fr/>. Consulté le 15 février 2021

<sup>4</sup> « Modèle de référence IFLA pour les bibliothèques », Wikipédia, [https://fr.wikipedia.org/wiki/Mod%C3%A8le\\_de\\_r%C3%A9f%C3%A9rence\\_IFLA\\_pour\\_les\\_biblioth%C3%A8ques](https://fr.wikipedia.org/wiki/Mod%C3%A8le_de_r%C3%A9f%C3%A9rence_IFLA_pour_les_biblioth%C3%A8ques). Consulté le 15 février 2021

<sup>5</sup> J'emprunte cette conception à Vincent Boulet, chef du service des Référentiels du département des Métadonnées de la BnF, dans sa conférence au colloque « Bibliographic Control in the Digital Age », le 9 février 2021, [https://www.youtube.com/watch?v=Z\\_MDPpzVR14](https://www.youtube.com/watch?v=Z_MDPpzVR14), vers 8'30. Consulté le 16 février 2021

Dans le premier cas, lorsque les éditeurs produisent des métadonnées, l'enjeu porte sur la qualité de ces métadonnées, qui est parfois médiocre, par exemple quand elles ne comportent pas d'identifiants uniques (ambiguïté), de métadonnées au niveau du chapitre d'un livre ou de l'article d'une revue<sup>6</sup> (granularité), quand certaines métadonnées manquent, comme les données de citations ou les résumés (incomplétude), ou ne sont pas ouvertes<sup>7</sup> (fermeture), quand les métadonnées ne sont pas lisibles par des machines (manque d'interopérabilité)... Les professionnels de l'information scientifique et technique (IST) ont sur ces points un rôle à jouer, notamment au niveau des agences bibliographiques, pour proposer par exemple des solutions de traitement en masse.

Mais les chercheurs et les bibliothèques de recherche se font également créateurs de métadonnées pour le contenu qu'ils créent et diffusent eux-mêmes notamment via des archives institutionnelles, ou dans des revues en open access éditées au sein des établissements. C'est le modèle de l'« *inside-out collection* » décrit par Lorcan Dempsey, vice-président d'OCLC, où la bibliothèque n'est plus seulement réceptrice de ressources et de métadonnées associées, mais participe au processus de création de ressources et donc des métadonnées qui les signalent<sup>8</sup>.

Dans le même temps, le web sémantique autorise une exposition de ces données à large échelle, au-delà même du monde des bibliothèques, dans une dimension plus collaborative, en facilitant les échanges de données, les alignements au moyen de formats et référentiels pivots. Ce travail aux échelles locale ou nationale peut ainsi se diffuser pour atteindre, hypothétiquement, à terme, l'idéal d'un index mondial référençant la production scientifique de façon exhaustive, ouverte et selon des standards de qualité élevés. Et ces standards de haute qualité ne peuvent être atteints que si la qualité existe dès le départ, car malgré les possibilités de traitements, de nettoyage, d'alignements, des données de mauvaise qualité à l'entrée produiront toujours des données de mauvaise qualité à la sortie<sup>9</sup>, avec même un risque de diffusion des erreurs dans le cadre des données liées.

L'investissement renforcé dans le signalement de la production locale n'a donc rien d'un repli sur cet échelon, car il s'accompagne d'un partage des données à l'échelle mondiale et d'une mutualisation. L'idée de contrôle bibliographique universel (CBU) s'est développée au sein de l'Ifla dans les années 1970. Selon elle, « chaque agence bibliographique nationale devait cataloguer les documents publiés dans son propre pays et établir les formes du nom des auteurs nationaux », pour que ces données soient « partagées et réutilisées dans le monde entier »<sup>10</sup>. Elle est toujours d'actualité, mais le web sémantique remet en débat cette notion, à travers l'éclosion de nouveaux standards, modèles conceptuels, identifiants, institutions, technologies<sup>11</sup>. Par analogie avec une sorte de CBU à l'heure du web sémantique, on peut voir les agences nationales et les bibliothèques comme investies d'une mission de gestion des données de leur périmètre, qui peut aller, dans le cas des productions scientifiques, jusqu'au niveau de

<sup>6</sup> « State of Open Monographs Series: Crossing the Rubicon – The Case for Making Chapters Visible », Jennifer Kemp et Mike Taylor, 2020, [www.digital-science.com/blog/news/state-of-open-monographs-series-making-chapters-visible/](https://www.digital-science.com/blog/news/state-of-open-monographs-series-making-chapters-visible/). Consulté le 9 février 2021

<sup>7</sup> « Crowdsourcing open citations with Croci. An analysis of the current status of open citations, and a proposal », Ivan Heibi, Silvio Peroni et David Shotton, 2019, <https://arxiv.org/pdf/1902.02534.pdf>. Consulté le 9 février 2021

<sup>8</sup> « Library Collections in the Life of the User: Two Directions », Lorcan Dempsey, 2016, *Liber Quarterly*, 26(4), <https://www.liberquarterly.eu/articles/10.18352/lq.10170/>. Consulté le 9 février 2021

<sup>9</sup> Selon la formule « Garbage in, garbage out », bien connue en informatique et au-delà. Voir « GIGO », Wikipédia, <https://fr.wikipedia.org/wiki/GIGO>. Consulté le 2 mars 2021

<sup>10</sup> « Déclaration professionnelle de l'Ifla sur le contrôle bibliographique universel », <https://www.ifla.org/FR/publications/node/92264>. Consulté le 25 janvier 2021

<sup>11</sup> Un récent colloque « Bibliographic Control in the Digital Ecosystem », tenu à Florence du 8 au 12 février 2021, témoigne de la centralité de ces évolutions : <https://www.bc2021.unifi.it/home>. Consulté le 15 février 2021

l'établissement universitaire dans le cadre de l'« *inside-out collection* » (identification des chercheurs et organismes locaux, signalement de leurs productions dans des archives ouvertes...). Pour les agences, le périmètre couvre notamment les éditeurs nationaux. L'enjeu est de parvenir à une très haute qualité de métadonnées, sur un périmètre propre relativement restreint qui rend possible la qualité parce qu'il est limité, et parce que les professionnels locaux sont les mieux placés pour savoir comment apporter à ces données la plus grande valeur ajoutée.

Le rôle des bibliothécaires et professionnels de l'IST apparaît alors : élaborer des métadonnées de haute qualité sur leur périmètre, local ou national. Un tel objectif requiert des compétences spécifiques car il nécessite le recours à des syntaxes et des vocabulaires communs, à des protocoles d'échange partagés, à des identifiants, des standards et des référentiels, mais aussi à des technologies de traitement des données. Il implique également une connaissance du cadre juridique et politique de la science ouverte, et de l'univers institutionnel du web sémantique. Ce sont ces différents éléments que nous nous efforcerons d'analyser dans ce mémoire, en examinant dans une première partie la notion de métadonnées et les principes et outils pour leur gestion (chapitre 1), puis le cadre institutionnel et politique de leur administration aux échelons national, européen et mondial (chapitre 2). Puis, dans une seconde partie, plus axée sur les pratiques des professionnels de l'IST dans les établissements, nous présenterons la question des identifiants pérennes (chapitre 3), puis la création de métadonnées en archives ouvertes et en édition open access (chapitre 4), et enfin les traitements automatisés permettant la gestion massive de données (chapitre 5).

# PARTIE A : LES MÉTADONNÉES DES RESSOURCES SCIENTIFIQUES ET LEUR ÉCOSYSTÈME

---

Dans un premier temps, nous nous attacherons à définir ce que sont les métadonnées dans leur diversité, et l'écosystème politique et économique dans lequel elles s'inscrivent, préalables nécessaires à une compréhension de l'action possible des professionnels de l'IST dans leur gestion.

## CHAPITRE 1 : UN ÉCOSYSTÈME COMPLEXE

### Des métadonnées diverses

Les métadonnées sont des données sur des données, permettant de décrire des ressources et d'en assurer la visibilité, par exemple dans une base de données, mais aussi la réutilisation, la conservation, etc. En conséquence de ces usages multiples, il existe différentes sortes de métadonnées, adaptées à un usage déterminé, dans un contexte donné (les métadonnées utiles aux éditeurs, aux chercheurs, aux financeurs, etc. ne sont pas nécessairement les mêmes).

On peut distinguer classiquement, en suivant par exemple Crossref<sup>12</sup>, les métadonnées bibliographiques, les métadonnées structurales et les métadonnées administratives.

- Les **métadonnées descriptives** ou bibliographiques permettent de décrire et citer une ressource. Il s'agit des noms ou des identifiants des auteurs, financeurs et structures impliquées, du titre, de la date de publication (ou de modification), d'identifiants comme l'ISSN ou l'ISBN, du DOI, du résumé, des mots-clés et vedettes-matières, etc.

- Les **métadonnées structurales** décrivent la structure de documents ou les relations entre des documents. Par exemple la structure d'une revue constituée de numéros successifs, d'un livre divisé en chapitre, ou des relations de citations entre différentes ressources, ou encore la relation entre différentes versions d'un même article. L'attribution de métadonnées à différents niveaux de granularité, par exemple au niveau de chaque chapitre par les éditeurs, constitue un enjeu.

- Les **métadonnées administratives** enfin renvoient aux informations relatives aux financements, aux licences d'utilisation (Creative Commons, etc.), le lien au texte complet, etc.

Le « Metadata Design and Best Practices » de l'université Cornell<sup>13</sup> propose deux catégories supplémentaires :

- Les **métadonnées de préservation** enregistrent des détails favorisant la conservation sur le long terme, comme l'historique des transformations (restructuration des fichiers, migrations vers de nouveaux formats, logiciels utilisés...).

- Les **métadonnées techniques** donnent des informations comme le format ou la

---

<sup>12</sup> « The wonderful world of metadata », Crossref, <https://www.crossref.org/education/metadata/>. Consulté le 26 janvier 2021

<sup>13</sup> « Metadata Design and Best Practices », Cornell University, <https://confluence.cornell.edu/display/culpublic/Metadata+Design+and+Best+Practices>. Consulté le 26 janvier 2021



taille de fichiers, etc.

Niso reprend ces diverses catégories<sup>14</sup>, en ajoutant les langages de balisage, qui permettent d'introduire des métadonnées au cœur même des textes.

Metadata Type	Example Properties	Primary Uses
Descriptive metadata	Title Author Subject Genre Publication date	Discovery Display Interoperability
Technical metadata	File type File size Creation date/time Compression scheme	Interoperability Digital object management Preservation
Preservation metadata	Checksum Preservation event	Interoperability Digital object management Preservation
Rights metadata	Copyright status License terms Rights holder	Interoperability Digital object management
Structural metadata	Sequence Place in hierarchy	Navigation
Markup languages	Paragraph Heading List Name Date	Navigation Interoperability

Tableau récapitulatif produit par Niso des divers types de métadonnées, de leurs propriétés et usages<sup>15</sup>

Une autre distinction importante peut être tracée entre métadonnées bibliographiques et données d'accès : les premiers éléments permettent de définir le contenu d'une œuvre rendant possible l'identification par l'utilisateur d'une ressource où il trouvera l'information qu'il recherche, les autres établissent une adresse physique ou numérique où trouver le document.

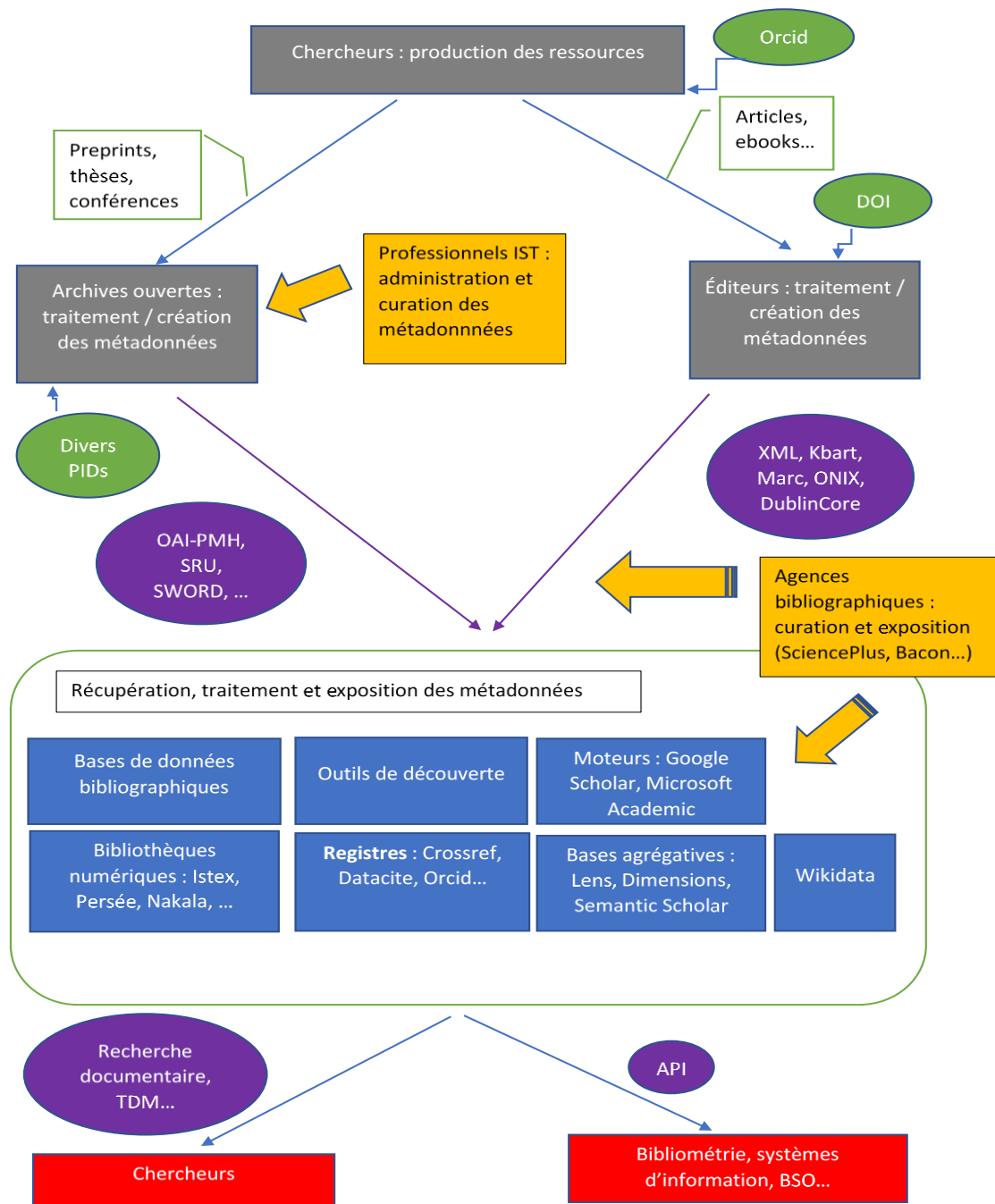
## Des acteurs et enjeux multiples dans un flux de métadonnées

Les métadonnées peuvent être pensées sous la forme d'un flux allant des producteurs aux utilisateurs, à travers différentes étapes de leur cycle de vie, où interviennent des acteurs divers. Ce « flux de métadonnées » peut à la fois rendre possibles une amélioration de la qualité des métadonnées au cours du processus (par enrichissements, nettoyage, corrections), mais aussi sa détérioration si les standards et protocoles employés ne permettent pas un échange optimal.

<sup>14</sup> « Understanding Metadata », Jenn Riley / Niso, 2017, [https://groups.niso.org/apps/group\\_public/download.php/17446/Understanding%20Metadata.pdf](https://groups.niso.org/apps/group_public/download.php/17446/Understanding%20Metadata.pdf). Consulté le 26 janvier 2021

<sup>15</sup> Ibid., p. 11

Schéma général des flux des métadonnées des productions scientifiques en format numérique



*De nombreux acteurs interviennent dans la chaîne de production des métadonnées de publications scientifiques, qui sont un matériau essentiel pour à la fois les identifier, les retrouver et y accéder, et qui servent aussi à des mesures de bibliométrie et de scientométrie.*

On peut distinguer, dans la « chaîne d’approvisionnement » (*supply chain*) des métadonnées des ressources scientifiques, six « parties prenantes » (*stakeholders*) en jeu<sup>16</sup> :

- Les **chercheurs** sont les producteurs des ressources, et en tant que tels ils sont aussi, en principe, à l’origine des métadonnées correspondantes. Pourtant, ils manquent de temps pour créer les métadonnées (d’autant plus que, faute d’échanges suffisants entre les systèmes, ils doivent souvent renseigner de multiples fois les mêmes informations, parfois selon des modèles différents suivant l’éditeur ou la plateforme...). Une certaine méconnaissance des enjeux des métadonnées, par exemple de l’importance des identifiants de chercheurs, est observée, liée à un mode de partage des ressources souvent plutôt informel (via les réseaux sociaux académiques, voire par courrier électronique). En conséquence, la qualité des métadonnées fournies par les chercheurs eux-mêmes est parfois sous-optimale. Une réflexion est sans doute à mener sur le partage des tâches dans l’administration des métadonnées entre chercheurs et professionnels de l’IST.

- Les **éditeurs scientifiques** utilisent des métadonnées dans leur chaîne de production, mais des contraintes de rentabilité pèsent parfois sur leur qualité. La maîtrise technique nécessaire, en constante évolution, est présente chez certains éditeurs, moins chez d’autres. Pourtant, ils voient majoritairement dans les métadonnées un champ à investir de façon prioritaire, dans la mesure où des métadonnées riches et de qualité permettent une augmentation des ventes. À l’heure actuelle, pourtant, les éditeurs sont encore loin de remplir un certain nombre d’exigences sur les métadonnées qu’ils diffusent. L’étude « The State of Journal Production and Access 2020 »<sup>17</sup> montre ainsi que si la plupart des éditeurs produisent des métadonnées de base (titre, auteurs, affiliations, résumé...) au niveau de l’article et lisibles par des machines, et plus de la moitié y incluent DOI, licence de réutilisation, et références, moins de 50% d’entre eux ajoutent d’autres identifiants pérennes (comme Orcid ou Grid) et des informations sur les financeurs. Il est cependant possible de mettre en place des initiatives proactives pour inciter les éditeurs à inclure dans leurs métadonnées des identifiants, comme le fait la British Library pour les Isni en collaboration avec plusieurs maisons d’édition<sup>18</sup>.

- Les **bibliothèques, ainsi que les agences bibliographiques (Abes et BnF en France, Library of Congress aux États-Unis, etc.) et les réseaux de bibliothèques qu’elles animent**, sont dans une position centrale dans le flux des métadonnées, entre les chercheurs, les éditeurs, les fournisseurs de services, d’un côté, et les utilisateurs finaux de ressources de l’autre. Ils sont ainsi dans une position centrale pour la gestion, l’enrichissement, la curation de ces métadonnées, afin d’améliorer leur qualité et leur interopérabilité.

- Les **dépôts de données** (archives ouvertes, entrepôts de données) sont au cœur du mouvement pour l’ouverture des productions scientifiques. Ils font face à la double injonction (dans une certaine mesure contradictoire) d’améliorer l’engagement des chercheurs (dans le cadre de l’auto-archivage) et d’améliorer la qualité des métadonnées. Souvent administrés localement (en dehors d’archives ouvertes nationales comme HAL ou de gros entrepôts généralistes comme Zenodo), ils peuvent se heurter à des problèmes

<sup>16</sup> « A literature review of scholarly communications metadata. Research Ideas and Outcomes », Gregg W. J., Erdmann C., Paglione LAD, Schneider J., Dean C., <https://doi.org/10.3897/rio.5.e38698>. Voir aussi, inspiré du précédent : « Communities », <http://www.metadata2020.org/communities/>. Consultés le 26 janvier 2021

<sup>17</sup> « The State of Journal Production and Access 2020 », Scholastica, <https://s3.amazonaws.com/marketing.scholasticahq.com/State-Journal-Production-Access-2020.pdf>. Consulté le 26 janvier 2021

<sup>18</sup> « Transitioning to the Next Generation of Metadata », Karen Smith-Yoshimura / OCLC, 2020, <https://www.oclc.org/content/dam/research/publications/2020/oclcresearch-transitioning-next-generation-metadata-a4.pdf>. Consulté le 26 janvier 2021

de moyens, de ressources humaines ou de compétences. C'est un autre enjeu très important des métadonnées pour la science ouverte sur lequel nous reviendrons.

- Les **fournisseurs de services** (plateformes, outils de découverte, résolveurs de liens) sont au cœur des problèmes d'interopérabilité, de manque de consistance des schémas de métadonnées (entre éditeurs et fournisseurs de services), de standards suivis par toute la communauté, d'identifiants largement reconnus et de problèmes de métadonnées périmées qui ne sont pas mises à jour par leurs créateurs<sup>19</sup>.

- Les **financeurs** enfin peuvent avoir une influence sur la manière dont les chercheurs qu'ils financent gèrent leurs données et leurs métadonnées. Les demandes en termes d'ouverture des publications et des données augmentent, et une plus grande attention à la qualité des données, avec par exemple des exigences de *data management plan* (plan de gestion de données) dans l'administration des projets de recherche financés.

### Focus : les métadonnées de ressources électroniques

Les ressources électroniques sont un bon exemple des problèmes qui peuvent se poser au sein de la chaîne de traitement des métadonnées entre les différentes parties prenantes évoquées ci-dessus. Les métadonnées portent tant sur des documents imprimés que sur des documents numériques, qui peuvent présenter le même contenu sur deux supports différents, mais le degré de maturité des métadonnées des documents imprimés et des documents électroniques est très différent, pour des raisons qui tiennent moins à la nature des documents qu'à leur mode d'acquisition et au volume de ces acquisitions.

En effet, les documents électroniques étant souvent achetés par les bibliothèques universitaires en masse, via des bouquets regroupant parfois des centaines de titres, le signalement dans les catalogues de bibliothèques est une tâche difficile et qui n'est à ce jour pas réalisée de façon exhaustive. Tandis que traditionnellement le catalogage des ressources papier est réalisé par des professionnels au titre à titre, cela n'est plus possible avec la documentation électronique. En conséquence, la qualité des métadonnées dépend de la qualité, hétérogène, des données fournies par les éditeurs : longs délais d'actualisation, liens rompus, absence de mots-clés, de résumés, d'identifiants, etc., sont fréquents et responsables d'expériences utilisateurs dégradées. Le récent rapport Couperin sur les pratiques documentaires des chercheurs est de ce point de vue éloquent : « Pour accéder au texte intégral, les chercheurs privilégient par ordre de priorité : Google Scholar, les sites des revues (via les éditeurs ou les agrégateurs, ces derniers très importants pour les SHS), les archives ouvertes, les réseaux sociaux et Sci-Hub. Les accès proposés par les bibliothèques sont perçus comme trop complexes et moins performants. »<sup>20</sup>

La question de ce que les professionnels de l'IST peuvent faire pour remédier à ces problèmes est donc centrale. Istex (25 millions de documents) est un exemple plutôt réussi de traitement de données des éditeurs (voir chapitre 5). Mais il s'agissait alors de traiter un stock de données et non de les traiter en flux continu, ce qui pose des problèmes

---

<sup>19</sup> « What Are the Common Causes of Full Text Linking Problems, and How Can Linking Be Improved? », ExLibris, 2017, [https://knowledge.exlibrisgroup.com/360\\_Services/360\\_Link/Knowledge\\_Articles/What\\_Are\\_the\\_Common\\_Causes\\_of\\_Full\\_Text\\_Linking\\_Problems%2C\\_and\\_How\\_Can\\_Linking\\_Be\\_Improved%3F](https://knowledge.exlibrisgroup.com/360_Services/360_Link/Knowledge_Articles/What_Are_the_Common_Causes_of_Full_Text_Linking_Problems%2C_and_How_Can_Linking_Be_Improved%3F). Consulté le 26 janvier 2021

<sup>20</sup> « Les pratiques de recherche documentaire des chercheurs français en 2020 : étude du consortium Couperin », Rapport Couperin n° 2, Marie-Pascale Baligand, Grégory Colcanap, Vincent Harnais, Françoise Rousseau-Hans, Christine Weil-Miko, Couperin.org, 2021, <https://hal.inrae.fr/hal-03148285/document>. Consulté le 2 mars 2021

supplémentaires : il est alors nécessaire d'aller chercher les données chez l'éditeur, et des problèmes de mise à jour des notices peuvent se poser a posteriori.

Il est donc important d'avoir des métadonnées de qualité en amont dans la chaîne de diffusion des métadonnées, de façon à ce que cette qualité se retrouve dans les différents outils où ces métadonnées de ressources électroniques sont exposées (catalogues, outils de découverte, moteurs de recherche...), ce qui exige une interopérabilité entre ces univers. Des bonnes pratiques, l'usage de protocoles normalisés peuvent permettre de répondre à cette exigence. Par exemple, KBart (« Knowledge Bases and Related Tools ») est une recommandation de Niso<sup>21</sup> conçue pour faciliter le transfert de métadonnées entre les fournisseurs de contenus électroniques (éditeurs...) et les bases de connaissances, permettant d'améliorer la découverte des ressources dans un catalogue de bibliothèque, et leur consultation via OpenURL<sup>22</sup>. Dans ce cadre, l'Abes propose le service Bacon (Base de connaissance nationale), qui « collecte, corrige, enrichit puis diffuse les métadonnées des ressources électroniques disponibles dans les établissements sous forme de fichiers KBart », tout en s'efforçant de collaborer en amont avec les éditeurs « en vue d'améliorer la qualité des métadonnées associées aux ressources composant leur offre éditoriale »<sup>23</sup>. Cela permet à l'Abes par exemple de maintenir des workflows d'import et de traitement de données d'Oxford University Press dans le Sudoc, en croisant les données des fichiers KBart et les ISBN des notices d'imprimés<sup>24</sup>.

## Les principes Fair : quatre piliers pour une bonne gestion des métadonnées

La qualité et l'interopérabilité des métadonnées exigent des bonnes pratiques tout au long de la chaîne de traitement de ces métadonnées. Une référence incontournable en matière de science ouverte sont les principes Fair, acronyme signifiant Findable, Accessible, Interoperable, Reusable (ou en français Facile à trouver, Accessible, Interopérable, Réutilisable) : ces quatre piliers sont les « lignes directrices pour améliorer la facilité de repérage, l'accessibilité, l'interopérabilité et la réutilisation des ressources numériques »<sup>25</sup>.

Ces principes ont été proposés dans un article de *Scientific Data* en 2017, « The Fair Guiding Principles for Scientific Data Management and Stewardship »<sup>26</sup>, partant du constat qu'une bonne gestion des données de la recherche à long terme permet la découverte et l'innovation, grâce à la transparence et la reproductibilité qui leur sont associées. Elles permettent aux chercheurs de voir leurs travaux mieux diffusés, cités et réutilisés, mais rendent aussi possible un traitement des données à grande échelle, notamment par des machines.

<sup>21</sup> « KBart: Knowledge Bases and Related Tools », Niso/UKSG KBart Working Group, 2010, <https://web.archive.org/web/20110716181454/http://www.niso.org/publications/rp/RP-2010-09.pdf>. Consulté le 10 février 2021

<sup>22</sup> « KBart Frequently Asked Questions », Niso, <https://www.niso.org/standards-committees/kbart/kbart-frequently-asked-questions>. Consulté le 10 février 2021

<sup>23</sup> « Bacon, Base de connaissance nationale », Abes, <https://abes.fr/documentation-electronique/bacon-signallement-ressources-electroniques/>. Consulté le 16 février 2021

<sup>24</sup> « Vers un nouveau workflow d'imports de données dans le Sudoc : les notices des ouvrages publiés par Oxford University Press », Punktokomo, 2018, <https://punktokomo.abes.fr/2018/11/08/import-courant-dans-le-sudoc-des-notices-imprimees-et-electroniques-des-ouvrages-publies-par-oxford-university-press-vers-un-nouveau-workflow-dimports/>. Consulté le 17 février 2021

<sup>25</sup> « Les principes Fair », Doranum, <https://doranum.fr/enjeux-benefices/principes-fair/>. Consulté le 26 janvier 2021

<sup>26</sup> « The Fair Guiding Principles for scientific data management and stewardship », Wilkinson, M., Dumontier, M., Aalbersberg, I. et al., *Sci Data* 3, 2016, <https://doi.org/10.1038/sdata.2016.18>. Consulté le 26 janvier 2021

Les principes Fair sont les suivants (dans une traduction de Doranum<sup>27</sup>) :

*Findable* : « Faciliter la découverte des données :

- Les données ont un PID (Persistent IDentifier ou identifiant pérenne en français)
- Les données sont décrites par des métadonnées
- Ces métadonnées spécifient le PID des données
- Les données sont déposées dans un entrepôt de données »

*Accessible* : « Permettre l'accès aux données et leur téléchargement :

- Les données sont accessibles à travers un protocole de communication standard
- Ce protocole est libre et ouvert
- Ce protocole permet un accès par authentification si besoin
- Les métadonnées restent accessibles même si les données ne le sont pas »

*Interoperable* : « Permettre l'exploitation des données quel que soit l'environnement informatique utilisé :

- Les données sont décrites avec un vocabulaire contrôlé
- Le vocabulaire utilisé respecte les principes Fair
- Les métadonnées sont reliées à d'autres données »

*Reusable* : « Permettre la réutilisation des données pour de futures recherches :

- Les métadonnées ont une pluralité d'attributs
- Une licence de réutilisation est attribuée aux données
- La description des données indique leur provenance
- Le partage des données suit les standards de la communauté scientifique »

Les principes Fair promeuvent donc des métadonnées riches, appuyées sur des standards (identifiants, protocoles, vocabulaires) et ouvertes, de façon à permettre leur exploitation par des humains comme par des machines. Soulignons cet aspect : la science ouverte ne peut se contenter que des articles ou des jeux de données soient déposés quelque part sur le web, sans qu'elles puissent être effectivement trouvées et récupérées par les utilisateurs potentiels. Des métadonnées accessibles librement et gratuitement sur une page web d'un éditeur, d'un chercheur ou via un moteur de recherche comme Google Scholar ne sont pas pour autant conformes aux principes Fair. C'est par la mise en place d'API permettant de récupérer les données automatiquement selon un protocole précis, que les données peuvent être véritablement qualifiées de Fair. Notons toutefois que, inversement, la conformité à ces principes n'implique pas automatiquement l'ouverture : les données doivent en effet être, selon l'expression consacrée, « aussi ouvertes que

---

<sup>27</sup> « Les principes Fair », Doranum, <https://doranum.fr/enjeux-benefices/principes-fair/>. Consulté le 26 janvier 2021

possible, aussi fermées que nécessaire »<sup>28</sup>. Néanmoins, cette précaution concerne surtout les données de la recherche, qui peuvent être sensibles (données personnelles, brevets, etc.), mais concernant les métadonnées, le principe d'ouverture ne peut souffrir que peu d'exceptions, même si les données sur lesquelles elles portent sont fermées.

## Un arsenal d'outils pour décrire les ressources scientifiques de façon standard

Les institutions documentaires peuvent, pour décrire les ressources qu'elles possèdent, diffusent ou produisent, utiliser une diversité de syntaxes, vocabulaires, thésaurus, schémas de métadonnées et ontologies<sup>29</sup>. Ces outils permettent de décrire des entités et les connaissances qu'on a sur elles, de façon organisée selon les exigences d'un champ scientifique ou de façon plus générale, et d'établir des liens entre ces entités.

La recherche d'interopérabilité et de partage commande de recourir préférentiellement à des standards, outils déjà existants, normalisés et utilisés à plus ou moins grande échelle, plutôt que d'en créer de nouveaux<sup>30</sup>, conformément aux principes Fair.

### Structures

D'abord, la standardisation concerne la structuration syntaxique du fichier contenant les métadonnées d'une ressource. Le langage XML (eXtensible Markup Language) permet d'organiser de manière hiérarchique des éléments à partir d'une racine, d'une façon qui soit lisible à la fois par des humains et par des machines. D'autres syntaxes, considérées comme moins complexes, comme Json ou Json-LD, sont également utilisées dans le monde des bibliothèques.

Dans le cadre du web sémantique et des données liées (*linked data*), le modèle RDF permet la structuration de l'information sous forme de triplets sujet-prédicat-objet, où chaque élément est identifié par une URI (Uniform Resource Identifier). Il permet ainsi d'exprimer à travers des *propriétés* (correspondant au prédicat) les relations entre des entités, elles-mêmes conçues comme des instances de *classes*.

### Schémas de métadonnées et ontologies

Le contenu sémantique remplissant ces structures est déterminé par une diversité de schémas de métadonnées, ontologies ou vocabulaires. Par exemple, on peut, au sein d'un fichier XML, avoir recours au vocabulaire Dublin Core, qui définit 15 métadonnées, facultatives et répétables, permettant de décrire de façon simple un document, notamment à des fins de moissonnage OAI-PMH. Ce Dublin Core Simple est complété par le standard

<sup>28</sup> À ce sujet, voir le rapport de Sparc Europe « Fair and Open Data », 2018, [file:///C:/Users/vince/Downloads/SPARCEurope\\_BriefingPaper\\_FAIROpenData.pdf](file:///C:/Users/vince/Downloads/SPARCEurope_BriefingPaper_FAIROpenData.pdf). Voir aussi la discussion « Open data et Fair : deux paradigmes différents ? », Team Open Data, 2017-2018, <https://teamopendata.org/t/open-data-et-fair-deux-paradigmes-differents/220>. Consultés le 24 février 2021

<sup>29</sup> Voir l'annexe sur les outils techniques pour plus de précisions sur les standards de métadonnées, les ontologies, les vocabulaires contrôlés et les protocoles d'échanges, ainsi que notre chapitre 3 dans la partie B pour une présentation plus approfondie de la question des identifiants pérennes.

<sup>30</sup> *Practical Ontologies for Information Professionals*, David Stuart, 2018, <https://doi.org/10.29085/9781783301522>. L'introduction de cet ouvrage peut être consultée ici : <https://pdfs.semanticscholar.org/2894/f4ae1e17b9bc04ada891c9c3eadbf7901d9d.pdf?ga=2.101863361.1745431819.1613472043-2010979089.1613472043>. Consultés le 16 février 2021

Dublin Core Terms<sup>31</sup>, qui propose une gamme plus large de métadonnées. Nous présentons ci-dessous un exemple d'une notice de métadonnées Dublin Core Simple extraite de HAL :

This XML file does not appear to have any style information associated with it. The document tree is shown below.

```
<?xml version="1.0" encoding="UTF-8" ?>
<oai_dc:dc xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/" xmlns:dc="http://purl.org/dc/elements/1.1/"
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xmlns:tei="http://www.tei-c.org/ns/1.0"
  xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/oai_dc/ http://www.openarchives.org/OAI/2.0/oai_dc.xsd
  http://purl.org/dc/elements/1.1/ http://dublincore.org/schemas/xmls/qdc/2008/02/11/dc.xsd">
  <dc:publisher>HAL CCSD</dc:publisher>
  <dc:title xml:lang="en">Vulnerability, resilience and adaptation of societies during major extreme storms during the Little
  Ice Age</dc:title>
  <dc:creator>Athimon, Emmanuelle</dc:creator>
  <dc:creator>Maanan, Mohamed</dc:creator>
  <dc:contributor>Littoral, Environnement, Télédétection, Géomatique UMR 6554 (LETG) ; Université de Caen Normandie (UNICAEN) ;
  Normandie Université (NU)-Normandie Université (NU)-Université d'Angers (UA)-Université de Nantes (UN)-École pratique des
  hautes études (EPHE) ; Université Paris sciences et lettres (PSL)-Université Paris sciences et lettres (PSL)-Université de
  Brest (UBO)-Université de Rennes 2 (UR2) ; Université de Rennes (UNIV-RENNES)-Université de Rennes (UNIV-RENNES)-Centre
  National de la Recherche Scientifique (CNRS)</dc:contributor>
  <dc:description>International audience</dc:description>
  <dc:source>ISSN: 1814-9324</dc:source>
  <dc:source>EISSN: 1814-9332</dc:source>
  <dc:source>Climate of the Past</dc:source>
  <dc:publisher>European Geosciences Union (EGU)</dc:publisher>
  <dc:identifiant>hal-03142758</dc:identifiant>
  <dc:identifiant>http://hal.univ-nantes.fr/hal-03142758</dc:identifiant>
  <dc:identifiant>http://hal.univ-nantes.fr/hal-03142758/document</dc:identifiant>
  <dc:identifiant>http://hal.univ-nantes.fr/hal-03142758/file/2018_cp-14-1487-2018.pdf</dc:identifiant>
  <dc:source>http://hal.univ-nantes.fr/hal-03142758</dc:source>
  <dc:source>Climate of the Past, European Geosciences Union (EGU), 2018, 14 (10), pp.1487-1497. &#x27E8;10.5194/cp-14-1487-
  2018&#x27E9;</dc:source>
  <dc:identifiant>DOI: 10.5194/cp-14-1487-2018</dc:identifiant>
  <dc:relation>info:eu-repo/semantics/altIdentifier/doi/10.5194/cp-14-1487-2018</dc:relation>
  <dc:language>en</dc:language>
  <dc:subject>[SHS.ENVI]Humanities and Social Sciences/Environmental studies</dc:subject>
  <dc:subject>[SDU.STU.CL]Sciences of the Universe [physics]/Earth Sciences/Climatology</dc:subject>
  <dc:type>info:eu-repo/semantics/article</dc:type>
  <dc:type>Journal articles</dc:type>
  <dc:description xml:lang="en"> This paper reviews more than 19 691 French historical documents from 14 French archive
  centres. To assess data from historical documents, a method has been applied that leads to a record of 101 extreme storms
  with damage, including 38 coastal floods. Thus, the results show periods of increasing and decreasing storm frequency. These
  periods are examined. Furthermore, coastal hazards have forced societies to adapt and develop specific skills, lifestyles and
  coping strategies. This paper analyses some responses of past societies to these hazards. By doing so, useful ideas may be
  (re)discovered by today's communities in order to enable us to adapt and develop resilience. Similarly, a thorough knowledge
  of past meteorological hazards may allow our societies to recreate a link with territory, particularly through the
  (re)construction of an effective memory of these phenomena. </dc:description>
  <dc:date>2018</dc:date>
  <dc:rights>info:eu-repo/semantics/OpenAccess</dc:rights>
</oai_dc:dc>
```

### Notice Dublin Core d'un article dans HAL<sup>32</sup>

Pour structurer le plein-texte lui-même et l'enrichir sémantiquement, on peut utiliser TEI (Text Encoding Initiative)<sup>33</sup>. Il permet par exemple de signifier qu'un mot est un nom de lieu, qu'il désigne Paris en France et non Paris au Texas, ou encore que tel texte est un poème, et ses parties des quatrains ou des sonnets. De nombreux éditeurs structurent les textes des articles selon ce langage de balisage, ce qui permet de disposer d'une ressource sémantiquement riche, par exemple à des fins de fouille de texte. Notons l'existence du logiciel Grobid qui permet de générer du plein-texte structuré en TEI automatiquement à partir de PDF<sup>34</sup>. Il a notamment été utilisé au sein du projet Istex<sup>35</sup>. Mais TEI peut aussi être utilisé pour encoder les métadonnées des ressources au sein du

<sup>31</sup> « DCMI Metadata Terms », Dublin Core, <https://www.dublincore.org/specifications/dublin-core/dcmi-terms/>. Consulté le 22 février 2021

<sup>32</sup> L'article correspondant est le suivant : « Vulnerability, resilience and adaptation of societies during major extreme storms during the Little Ice Age », Emmanuelle Athimon et Mohamed Maanan, 2021, <https://hal.univ-brest.fr/hal-03142758v1>. Consulté le 16 février 2021

<sup>33</sup> « Text Encoding Initiative », Wikipédia, [https://fr.wikipedia.org/wiki/Text\\_Encoding\\_Initiative](https://fr.wikipedia.org/wiki/Text_Encoding_Initiative). Consulté le 1<sup>er</sup> février 2021

<sup>34</sup> « Grobid - Information Extraction from Scientific Publications », Laurent Romary et Patrice Lopez, 2015, <https://hal.inria.fr/hal-01673305/document>. Consulté le 21 février 2021

<sup>35</sup> « Fulltexts structurés à partir des PDFs avec Grobid », Blog Istex, 2017, <https://blog.istex.fr/fulltexts-structures-a-partir-des-pdfs-avec-grobid/>. Consulté le 21 février 2021



module <teiHeader>, avec des schémas propres à différents projets, car TEI est adapté à la granularité des données et offre la plasticité nécessaire, ce qui en retour expose à des faiblesses d'interopérabilité<sup>36</sup>.

Des vocabulaires comme RDA<sup>37</sup>, Mods<sup>38</sup>, Mads<sup>39</sup>, Mets<sup>40</sup>, Jats<sup>41</sup>, Bibo<sup>42</sup>, Cito<sup>43</sup>, Foaf<sup>44</sup> sont aussi utilisés pour décrire des ressources documentaires, avec chacun des objectifs et des propriétés spécifiques<sup>45</sup>. Par exemple, Mads (Metadata Authority Description Schema) modélise les données d'autorité, Jats (Journal Article Tag Suite) permet spécifiquement la description d'articles scientifiques, Foaf (Friend of a friend) permet de décrire les personnes et leurs liens, Bibo (Bibliographic Ontology) permet la description de ressources bibliographiques sur le web sémantique, Cito (Citation Typing Ontology) décrit des relations de citation entre ressources et rend possible de distinguer entre plusieurs types de citations (avec des propriétés comme <cito:citesAsDataSource> ou <cito:disagreeWith>...) <sup>46</sup>.

Au-delà du monde des bibliothèques, Schema.org est une initiative de moteurs de recherche commerciaux et autres entreprises du numériques (lancée en 2011 par Google, Microsoft, Yahoo and Yandex) pour structurer les métadonnées d'un site web, appartenant à la famille des microdonnées<sup>47</sup>. Il s'agit d'informations lisibles par les robots d'indexation des moteurs de recherche, formulées dans un vocabulaire hiérarchisé constitué de « 829 types, 1351 propriétés et 339 valeurs d'énumérations »<sup>48</sup>.

Des ontologies de de plus haut niveau comme OWL<sup>49</sup>, Skos<sup>50</sup>, RDFS<sup>51</sup>, permettent de représenter les vocabulaires eux-mêmes, en apportant les outils pour identifier les classes d'entités, les propriétés, et les relations entre ces entités (relations d'inclusion, d'exclusion, de transitivité, de réciprocité, etc.)<sup>52</sup>. Là encore, chacun dispose de

<sup>36</sup> « Best Practices for TEI in Libraries », Text Encoding Initiative, <https://tei-c.org/extra/teilibraries/4.0.0/bptl-driver.html#header-content>. Consulté le 25 février 2021

<sup>37</sup> « Ressources : description et accès », Wikipédia, [https://fr.wikipedia.org/wiki/Ressources:\\_description\\_et\\_acc%C3%A8s](https://fr.wikipedia.org/wiki/Ressources:_description_et_acc%C3%A8s). Consulté le 21 février 2021

<sup>38</sup> « Mods RDF Ontology », 2013, <https://www.loc.gov/standards/mods/modsrdf/primer.html#namespaces>. Consulté le 21 février 2021

<sup>39</sup> « Mads/RDF (Metadata Authority Description Schema in RDF) », <https://id.loc.gov/ontologies/madsrdf/v1.html#>. Consulté le 21 février 2021

<sup>40</sup> « Mets : Metadata Encoding and Transmission Standard », BnF, <https://www.bnf.fr/fr/mets-metadata-encoding-and-transmission-standard>. Consulté le 21 février 2021

<sup>41</sup> « Journal Article Tag Suite », Wikipédia, [https://en.wikipedia.org/wiki/Journal\\_Article\\_Tag\\_Suite](https://en.wikipedia.org/wiki/Journal_Article_Tag_Suite). Consulté le 21 février 2021

<sup>42</sup> « Bibliographic Ontology Specification », Bibliontology, 2009, <http://bibliontology.com/>. Consulté le 21 février 2021

<sup>43</sup> « Cito, the Citation Typing Ontology », Github, <https://sparontologies.github.io/cito/current/cito.html>. Consulté le 22 février 2021

<sup>44</sup> « Foaf Vocabulary Specification 0.99 », 2014, <http://xmlns.com/foaf/spec/>. Consulté le 21 février 2021

<sup>45</sup> Pour déterminer les outils (schémas, ontologies...) pertinents pour les besoins d'un établissement ou d'un projet particulier, on peut signaler l'outil Fairsharing (<https://fairsharing.org/>), qui recense les standards, mais aussi les entrepôts, institutions et politiques conformes aux principes Fair, et partagés dans des communautés plus ou moins larges. L'outil peut être utilisé dans les établissements, dans le cadre de la recherche de modèles de métadonnées et d'entrepôts disciplinaires pour exposer des données. On peut aussi mentionner LOV (Linked Open Vocabularies, <https://lov.linkeddata.es/dataset/lov/>), qui recense plus de 700 vocabulaires du web sémantique, mais permet aussi de rechercher directement les termes (classes et propriétés) de ces vocabulaires.

<sup>46</sup> Par exemple, la plateforme Persée « utilise, entre autres, les vocabulaires “foaf” qui définit les relations entre personnes, “bibo” qui définit les relations entre entités bibliographiques, “dcterms” (Dublin core) qui définit un jeu de métadonnées classiques, “cito” qui définit les liens entre documents... » Voir « Qu'est-ce qu'un triplestore ? », Data Persée, <http://data.persée.fr/ressources/quest-ce-quun-triplestore/>. Consulté le 26 janvier 2021

<sup>47</sup> « Microdonnée », Wikipédia, <https://fr.wikipedia.org/wiki/Microdonn%C3%A9e>. Consulté le 9 février 2021

<sup>48</sup> « Organization of schemas », Schema.org, <https://schema.org/docs/schemas.html>. Consulté le 9 février 2021

<sup>49</sup> « OWL Web Ontology Language Overview », W3C, 2004, <https://www.w3.org/TR/owl-features/>. Consulté le 21 février 2021

<sup>50</sup> « SKOS Simple Knowledge Organization System Primer », W3C, 2009, <https://www.w3.org/TR/2009/NOTE-skos-primer-20090818/>. Consulté le 21 février 2021

<sup>51</sup> « RDF Schema 1.1 », W3C, <https://www.w3.org/TR/rdf-schema/>. Consulté le 22 février 2021

<sup>52</sup> « What is the difference between RDF and OWL? », <https://stackoverflow.com/questions/1740341/what-is-the-difference-between-rdf-and-owl>. Consulté le 24 février 2021

fonctionnalités parmi lesquelles les utilisateurs peuvent puiser selon leurs besoins. Par exemple, Skos permet spécifiquement de « représenter des thésaurus documentaires, classifications ou d'autres types de vocabulaires contrôlés ou de langages documentaires »<sup>53</sup>, à l'aide d'attributs comme `skos:Concept`, `skos:prefLabel`, `skos:altLabel`, `skos:hiddenLabel`, `skos:broader` et `skos:narrower`, `skos:related`, etc. Les vocabulaires Rameau et LCSH ont été publiés en Skos<sup>54</sup>, ce qui permet d'aligner les deux thésaurus.

Pour s'adapter finement à des besoins spécifiques auxquels un standard particulier ne répond pas à lui seul, il est possible de combiner différents vocabulaires pour décrire des ressources. Un standard de métadonnées ou un profil d'application est choisi en fonction de son utilisation, de la nature de la ressource (vidéo, image, jeu de données, article...), du domaine disciplinaire, de la communauté concernée<sup>55</sup>. On peut en effet avoir recours à des standards interdisciplinaires comme Dublin Core, très généraux et peu spécifiques, mais aussi à des standards disciplinaires permettant de décrire plus finement des ressources, par exemple en écologie l'Ecological Metadata Language<sup>56</sup>.

Les différents entrepôts de données, archives, etc., peuvent aussi avoir des exigences particulières en matière de standards de métadonnées. Ils peuvent pour cela définir des « profils d'application »<sup>57</sup>, combinant des métadonnées de différents vocabulaires. Ainsi, le profil d'application d'Openaire mêle des éléments issus des espaces de noms Dublin Core, DCTerms, DataCite Metadata Schema et des éléments « maisons »<sup>58</sup>.

Des thésaurus, tel Rameau en français, LCSH pour l'univers anglo-saxon, ou encore MeSH pour le domaine médical (et sa version française fMeSH) et de nombreux autres vocabulaires spécialisés, sont utilisés pour l'indexation matière des ressources documentaires. Un répertoire de vedettes-matières comme Rameau fait cependant l'objet d'une réforme (en cours) pour l'adapter au web sémantique<sup>59</sup>. Des technologies sont à présent disponibles pour produire de l'indexation matière automatiquement, soit en récupérant des mots-clés dans d'autres bases de données (par exemple, en s'appuyant sur le modèle Ifla-LRM, à partir d'autres Manifestations de la même Œuvre), soit par des procédures d'extraction de concepts dans le document lui-même, au moyen de techniques de traitement automatique de la langue ou d'apprentissage machine<sup>60</sup>.

## Licences

Les principes Fair exigent enfin de définir les conditions de réutilisation des données et des métadonnées, ce qui implique de diffuser ces données sous une licence qui

<sup>53</sup> « Simple Knowledge Organization System », Wikipédia, [https://fr.wikipedia.org/wiki/Simple\\_Knowledge\\_Organization\\_System](https://fr.wikipedia.org/wiki/Simple_Knowledge_Organization_System). Consulté le 10 février 2021

<sup>54</sup> « Rameau et Skos », Antoine Isaac et Thierry Bouchet, *Arabesques* n° 54, <https://publications-prairial.fr/arabesques/index.php?id=2109>. Consulté le 10 février 2021

<sup>55</sup> « Les standards de métadonnées : pourquoi et lequel ? », Doranum, <https://doranum.fr/metadonnees-standards-formats/standard-metadonnees/>. Consulté le 26 janvier 2021

<sup>56</sup> « EML; Ecological Metadata Language » FairSharing, <https://doi.org/10.25504/FAIRsharing.r3vtvx>. Consulté le 17 février 2021

<sup>57</sup> « Profil d'application », Wikipédia, [https://fr.wikipedia.org/wiki/Profil\\_d%27application](https://fr.wikipedia.org/wiki/Profil_d%27application). Consulté le 15 février 2021

<sup>58</sup> « Application Profile Overview », Openaire, [https://openaire-guidelines-for-literature-repository-managers.readthedocs.io/en/v4.0.0/application\\_profile.html#application-profile](https://openaire-guidelines-for-literature-repository-managers.readthedocs.io/en/v4.0.0/application_profile.html#application-profile). Consulté le 3 février 2021

<sup>59</sup> « Réformer Rameau », BnF, <https://rameau.bnf.fr/syntaxe/reformer>. Consulté le 22 février 2021. Voir aussi *L'indexation matière en transition. De la réforme de Rameau à l'indexation automatique*, Étienne Cavalie, 2019, <https://www.decitre.fr/livres/l-indexation-matiere-en-transition-9782765416234.html>. Consulté le 22 février 2021

<sup>60</sup> Voir à ce sujet « Indexation automatique des documents : de nouvelles métadonnées pour de nouveaux services », de Jean-Philippe Moreux, dans *L'indexation matière en transition*, op. cit.

en autorise le partage et la réutilisation, selon des conditions précises. Les plus utilisées pour les métadonnées, en France, sont les suivantes :

- Les licences Creative Commons permettent d'ouvrir les données tout en définissant « plusieurs restrictions, comme l'interdiction d'usage commercial ou de modification. »<sup>61</sup> Pour les métadonnées, la licence CC0 peut être utilisée pour indiquer qu'aucun droit d'auteur ne s'applique (proche du domaine public). C'est ce que fait l'Abes pour les données de Bacon par exemple<sup>62</sup>, ainsi que Wikidata ou encore Europeana.

- La Licence ouverte Etalab « a été conçue par le gouvernement français pour faciliter la mise en place de l'open data. Elle équivaut à la licence CC-BY. »<sup>63</sup> Elle ne s'applique qu'aux données publiques, en l'absence de droits patrimoniaux. Les données du Sudoc, d'IdRef, de Theses.fr, de Calames, sont diffusées sous cette licence, tout comme les métadonnées descriptives de la BnF ou encore les données de ScanR.

### Bilan d'étape

Dans ce premier chapitre, nous avons vu que l'écosystème des métadonnées des publications scientifiques et de recherche impliquait des acteurs multiples, au sein d'un flux où les bibliothèques occupent une place centrale, entre les producteurs et les utilisateurs. Nous avons aussi vu que la fluidité de ce flux était conditionnée par des principes, des bonnes pratiques et par l'utilisation d'outils et de techniques appropriés, que nous avons brièvement présentés (standards, identifiants, licences...). La question que nous poserons dans la suite de ce mémoire est celle de savoir comment les professionnels de l'IST peuvent œuvrer à une meilleure qualité des métadonnées en intervenant dans ce flux, au moyen des principes et outils mentionnés. Pour pouvoir faire un usage approprié de ces outils, une connaissance des institutions qui les gèrent, ainsi que de la gouvernance de la science ouverte qui modèlent l'environnement (politique, économique et juridique) où ils peuvent se déployer au mieux, est nécessaire. Ce sera l'objet du deuxième chapitre.

---

<sup>61</sup> « Les principes Fair », Doranum, <https://doranum.fr/enjeux-benefices/principes-fair/>. Consulté le 26 janvier 2021

<sup>62</sup> « Métadonnées libres pour le signalement de la documentation électronique », Abes, <https://bacon.abes.fr/>. Consulté le 26 janvier 2021

<sup>63</sup> « Les principes Fair », Doranum, op. cit.

## CHAPITRE 2 : GOUVERNANCE DE LA SCIENCE OUVERTE ET DES MÉTADONNÉES DE LA RECHERCHE

Dans ce chapitre, nous allons donner les grandes lignes de l'organisation institutionnelle de la science ouverte et des métadonnées ouvertes à différentes échelles, nationale, européenne et mondiale.

### **En France : un cadre incitatif à l'ouverture des données et un écosystème d'acteurs**

#### *La loi pour une République numérique et le plan national pour la science ouverte*

L'accès ouvert aux publications scientifiques est rendu possible en France par la loi pour une République numérique<sup>64</sup> de 2016, qui autorise les articles financés à au moins 50% sur fonds publics à être rendus accessibles en accès ouvert y compris quand ils sont publiés par ailleurs dans une revue, avec un délai de 6 mois maximum pour les STM et un an pour les SHS. Cela n'a pas cependant de caractère obligatoire.

Le Plan national pour la science ouverte<sup>65</sup> structure en France les efforts pour le développement de la science ouverte. Lancé par le Mesri en 2018, il est déployé autour de trois axes : généraliser l'accès ouvert aux publications ; structurer et ouvrir les données de la recherche ; s'inscrire dans une dynamique durable, européenne et internationale. Ces éléments ont des implications pour la gestion des métadonnées de la recherche, qui se trouvent ne pas être directement mentionnées dans le plan, quoique indispensables.

Le premier axe veut rendre obligatoire la mise à disposition en accès ouvert des recherches financées sur appels à projets sur fonds publics, notamment dans HAL, dont le rôle central est ainsi affirmé. Le Plan prévoit également le soutien à l'édition en accès ouvert. Ces dimensions impliquent un travail sur les métadonnées et le signalement afin que ces productions soient aussi faciles à trouver (pour reprendre la terminologie Fair) dans les bases de données internationales que les publications en accès fermé. En outre, le plan indique vouloir « simplifier le dépôt par les chercheurs qui publient en accès ouvert sur d'autres plateformes dans le monde », ce qui a des implications en termes d'interopérabilité pour éviter des saisies multiples.

Le deuxième axe implique que « les données produites par la recherche publique française soient progressivement structurées en conformité avec les principes Fair » mentionnés dans le chapitre 1. Il propose le recours à des « entrepôts de données certifiés » et la généralisation des plans de gestion de données.

Le troisième axe propose la participation à des initiatives transnationales comme Research Data Alliance (« réseau international définissant les bonnes pratiques dans le domaine des données de la recherche »), l'EOSC (European Open Science Cloud), Go Fair, Openaire, I4OC, le DOAJ (Directory of Open Access Journals), Operas, Scoss... Il

---

<sup>64</sup> « La loi République numérique », Science ouverte France, <https://scienceouverte.couperin.org/la-loi-numerique/>. Consulté le 26 janvier 2021

<sup>65</sup> « Plan national pour la science ouverte », Comité pour la science ouverte, 2018, <https://www.ouvrirelascience.fr/plan-national-pour-la-science-ouverte/>. Consulté le 26 janvier 2021

est aussi proposé d'« adhérer au niveau national à Orcid » et de « contribuer » à Crossref et Datacite, pour ce qui concerne plus étroitement les métadonnées, en l'occurrence les identifiants de chercheurs et de publications.

Il était également proposé de créer un baromètre de la science ouverte (BSO) ainsi que d'« enrichir ScanR, moteur de la recherche et de l'innovation, et Isidore, plateforme de recherche permettant l'accès aux données numériques des sciences humaines et sociales ».

Enfin, à la toute fin de l'année 2020, le rapport Bothorel « *pour une politique publique de la donnée* »<sup>66</sup> a fait un certain nombre de recommandations sur la gestion des données publiques pour inciter les acteurs publics à se lancer vraiment sur la voie de l'open data. Cela inclut les données de l'enseignement supérieur et de la recherche, mais cela va bien au-delà et permet de constater les réticences dans tous les secteurs publics vis-à-vis de l'ouverture des données, et le manque de mutualisation qui en est la conséquence. Nous aurons l'occasion de voir ce constat se matérialiser dans le secteur de la recherche dans la suite de ce document.

Ces initiatives diverses n'ont pas de valeur contraignante, y compris la loi pour une République numérique, qui autorise sans imposer le dépôt en archives ouvertes, avec un embargo. Pour autant, elles peuvent servir de socle à des politiques d'établissements qui choisiraient d'adopter des mandats de dépôt par exemple. Elles peuvent aussi être des points d'appui lors de négociations avec les éditeurs pour la mise à disposition de leurs métadonnées. Elles replacent aussi la France dans un écosystème international soutenant l'ouverture des données, par exemple avec le soutien à I4OC ou l'adhésion nationale à Orcid. Le PNSO a en outre mis en place le Fonds national pour la science ouverte, qui finance des projets parfois pionniers pour le développement de métadonnées et de données de qualité<sup>67</sup>, ainsi que le Baromètre de la science ouverte, qui est à la fois un projet intéressant de traitement des données de l'ESR et un indicateur sur l'état de la science ouverte à l'échelle nationale (nous reviendrons largement sur ce projet dans la partie 5). Ainsi, le PNSO met en place à la fois des incitations au développement de l'accès ouvert, ce qui occasionne un volume plus important de métadonnées à traiter (notamment dans les archives ouvertes), et des outils pour améliorer le traitement de ces données, facilitant cette gestion de la masse.

### *Les agences bibliographiques : l'Abes et la BnF*

La France compte deux agences bibliographiques : la BnF, Bibliothèque nationale de France, et l'Abes, Agence bibliographique de l'enseignement supérieur. À ce titre, selon l'Ifla, elles sont en charge de la création de métadonnées pour les productions éditoriales nationales, ainsi que l'alimentation de référentiels d'autorités<sup>68</sup>.

En conséquence, **la BnF** a « pour responsabilité d'établir des données d'autorité de référence pour les entités du domaine national »<sup>69</sup>. Ce travail est notamment mené par le

---

<sup>66</sup> « Pour une politique publique de la donnée, des algorithmes et des codes sources », Mission Bothorel, 2020, [https://www.gouvernement.fr/sites/default/files/contenu/piece-jointe/2020/12/rapport\\_-\\_pour\\_une\\_politique\\_publique\\_de\\_la\\_donnee\\_-\\_23.12.2020\\_0.pdf](https://www.gouvernement.fr/sites/default/files/contenu/piece-jointe/2020/12/rapport_-_pour_une_politique_publique_de_la_donnee_-_23.12.2020_0.pdf). Consulté le 26 janvier 2021

<sup>67</sup> « Résultats du premier appel à projet du Fonds national pour la science ouverte en faveur de l'édition scientifique ouverte », Mesri, 2020, <https://www.enseignementsup-recherche.gouv.fr/cid155105/resultats-premier-appel-projet-fonds-national-pour-science-ouverte-faveur-edition-scientifique-ouverte.html>. Consulté le 24 février 2021

<sup>68</sup> « Ifla Professional Statement on Universal Bibliographic Control », Ifla, <https://www.ifla.org/files/assets/bibliography/Documents/ifla-professional-statement-on-ubc-en.pdf>. Consulté le 26 janvier 2021

<sup>69</sup> « Données d'autorité à la BnF », Bibliothèque nationale de France, <https://www.bnf.fr/fr/donnees-autorite-bnf>. Consulté le 26 janvier 2021

département des Métadonnées<sup>70</sup>. Son travail quotidien dans ce cadre est d'appliquer « aux données d'autorité un processus de contrôle qualité ». Elle participe aussi à ce titre au Vif (Virtual International Authority File), elle est l'agence d'enregistrement Isni (International Standard Name Identifier) dans le cadre du dépôt légal, et héberge le centre national Rameau (Répertoire d'autorité-matière encyclopédique et alphabétique unifié). Elle assure des missions de diffusion de ces données (y compris sur le web sémantique via data.bnf.fr), et de formation en direction de nombreux agents de l'information bibliographique en France<sup>71</sup>. Elle pilote avec l'Abes le programme de Transition bibliographique<sup>72</sup>, reposant sur un nouveau code de catalogage par entités (selon le modèle Ifla-LRM) rendant possible l'exposition des catalogues de bibliothèques sur le web de données. Dans ce cadre, elle met en place actuellement Noemi (Nouer les Œuvres, les Expressions, les Manifestations et les Items)<sup>73</sup>, une nouvelle application de catalogage. Elle vise à « positionner le rôle des catalogueurs de plus en plus sur des actions d'enrichissement de la notice, de création de liens entre entités »<sup>74</sup>.

Des initiatives de traitement de données en masse sont menées par la BnF, notamment de clusterisation des auteurs et des documents, permettant de créer des notices d'œuvres dans data.bnf.fr<sup>75</sup>. Des projets d'indexation automatique sont aussi en cours, par exemple pour étendre l'indexation d'une Expression d'une Œuvre à l'ensemble des Expressions de cette même œuvre, grâce aux liens créés dans le cadre du modèle Ifla-LRM.

L'Abes, de son côté, « met à la disposition des professionnels de l'IST différents outils et services documentaires au bénéfice de l'enseignement supérieur et de la recherche »<sup>76</sup>. Son périmètre d'activité est donc plus centré sur ce qui intéresse notre sujet, les publications scientifiques.

Parmi les services mis en place par l'Abes qui nous intéressent dans le cadre de cette recherche, on trouve notamment différents catalogues, au premier rang desquels le Sudoc.fr, catalogue collectif de l'ESR, qui recense les collections de 163 établissements documentaires représentant 1 536 bibliothèques académiques<sup>77</sup>, ainsi que les publications en série de près de 3 000 bibliothèques du réseau Sudoc-PS, dont quelque 1 500 bibliothèques académiques<sup>78</sup>. Mais aussi d'autres outils, catalogues ou référentiels :

- Calames<sup>79</sup> : catalogue des manuscrits et archives de l'enseignement supérieur

---

<sup>70</sup> Le département des métadonnées compte 65 agents répartis en trois services correspondant au cycle de vie des données à la BnF, soit à l'entrée et la structuration des données, à l'enrichissement, puis à la diffusion : le service ingénierie des métadonnées (data.bnf.fr, corrections de masse, traitements automatisés, chargement de données, formats...) ; le service référentiels (données d'autorités, normalisation, et centre national Rameau) ; le service diffusion des métadonnées (exploitation et récupération des données, catalogues, centre ISSN, publics empêchés). Le département a une mission interne à la BnF (cohérence des catalogues, coordination et formation des catalogueurs notamment) et une mission d'agence bibliographique à l'échelle nationale, avec un partage des tâches avec l'Abes.

<sup>71</sup> « Quelles sont actuellement les missions de l'ABN (agence bibliographique nationale) au sein de la BnF ? », Enssib, <https://www.enssib.fr/services-et-ressources/questions-reponses/quelles-sont-actuellement-les-missions-de-labn-agence>

<sup>72</sup> Transition bibliographique, <https://www.transition-bibliographique.fr/>. Consulté le 26 janvier 2021

<sup>73</sup> « Noemi : vers un nouvel outil de production des métadonnées de la BnF », Bibliothèque nationale de France, <https://www.bnf.fr/fr/noemi-vers-un-nouvel-outil-de-production-des-metadonnees-de-la-bnf>. Consulté le 26 janvier 2021

<sup>74</sup> Ibid.

<sup>75</sup> « Processus expérimental : les œuvres générées automatiquement », Data BnF, <https://data.bnf.fr/fr/data-enrichment>. Consulté le 26 janvier 2021

<sup>76</sup> « Au cœur des missions de l'Abes », Abes, <http://www.abes.fr/Connaitre-l-Abes/Missions>. Consulté le 26 janvier 2021

<sup>77</sup> « Le Sudoc, réseau national de signalement des données bibliographiques pour l'ESR », Abes, 2020, <https://abes.fr/reseau-sudoc/le-reseau/etablissements-membres/>. Consulté le 17 février 2021

<sup>78</sup> « Les établissements membres du Réseau Sudoc PS », Abes, <https://abes.fr/reseau-sudoc-ps/le-reseau/etablissements-sudoc-ps/>. Consulté le 5 mars 2020

<sup>79</sup> Calames, Abes, <http://www.calames.abes.fr/pub/>. Consulté le 26 janvier 2021

- IdRef<sup>80</sup> : référentiel des autorités Sudoc
- Theses.fr<sup>81</sup> : moteur de recherche des thèses françaises
- Bacon<sup>82</sup> : base de connaissance nationale, qui facilite le signalement de la documentation électronique.

De plus, l'Abes assume certaines missions nationales :

- la normalisation documentaire, notamment de l'Unimarc, et le co-pilotage avec la Bnf du programme de Transition bibliographique ;
- la conversion rétrospective des catalogues ;
- le signalement des thèses.

Pour le signalement des thèses, l'Abes a développé deux applications : Star pour l'archivage électronique des thèses, qui dispose d'un schéma de métadonnées TEF spécifique aux thèses<sup>83</sup>, ensuite converti dans le Sudoc en Unimarc, en Dublin Core dans le réservoir OAI-PMH, et en TEI dans l'archive TEL<sup>84</sup> ; et Step<sup>85</sup>, pour le référencement des thèses en préparation.

En lien avec le référentiel d'autorités IdRef, l'Abes développe des outils de curation des métadonnées comme Paprika, interface qui permet le contrôle qualité des liens entre notices d'autorité et notices bibliographiques, incluant Qualinka<sup>86</sup>, application d'intelligence artificielle qui évalue automatiquement la qualité des liens entre points d'accès et autorités personnes<sup>87</sup>.

Les deux agences BnF et Abes copilotent le projet de Fichier national d'entités (FNE), projet lancé en 2017, et encore en cours d'élaboration (lancement prévu en 2023), qui a pour objectif la mise en place d'une plateforme de coproduction d'entités sous licence Etalab, associant les réseaux de l'Abes et de la BnF, ainsi que des institutions culturelles<sup>88</sup>. Ses objectifs sont de « mieux intégrer les référentiels produits par les institutions culturelles et scientifiques dans le web de données et en faciliter la réutilisation », de « mutualiser la production des données pour en améliorer la qualité et la fiabilité, en s'appuyant sur des outils modernisés » et de « développer dans la communauté professionnelle une stratégie partagée sur les données basée sur des bonnes pratiques communes ».

Le FNE, d'abord alimenté par les réservoirs d'autorités de l'Abes et de la BnF, a vocation à terme à agréger d'autres jeux de données, comme des « référentiels locaux des collectivités territoriales ou issus de la recherche et des institutions culturelles, autorités archivistiques... »<sup>89</sup>. Il n'est également pas exclu « de compléter les données présentes

<sup>80</sup> IdRef, Abes, <https://www.idref.fr/>. Consulté le 26 janvier 2021

<sup>81</sup> Theses.fr, Abes, <http://theses.fr/>. Consulté le 26 janvier 2021

<sup>82</sup> « Pourquoi Bacon ? », Abes, [https://bacon.abes.fr/pourquoi\\_bacon.html](https://bacon.abes.fr/pourquoi_bacon.html). Consulté le 26 janvier 2021

<sup>83</sup> « Les métadonnées des thèses électroniques françaises », Theses.fr, <https://www.theses.fr/schemas/tef/recommandation/index.html>. Consulté le 26 janvier 2021

<sup>84</sup> « Le signalement des thèses de doctorat », Abes, <http://www.abes.fr/Theses/Applications-pour-le-signalement-des-theses/Star-Signalement-des-Theses-ARchivage>. Consulté le 26 janvier 2021

<sup>85</sup> « Le signalement des thèses de doctorat », Abes, <http://www.abes.fr/Theses/Applications-pour-le-signalement-des-theses/Star-Signalement-des-Theses-ARchivage>. Consulté le 26 janvier 2021

<sup>86</sup> « IdRef, Paprika and Qualinka. A toolbox for authority data quality and interoperability », Aline Le Provost et Yann Nicolas, <https://hal.archives-ouvertes.fr/hal-02563630/document>. Consulté le 26 janvier 2021

<sup>87</sup> « La curation, un enjeu pour la gestion des données numériques », Aline Le Provost, *Arabesques* n° 97, 2020, <https://publications-prairial.fr/arabesques/index.php?id=1793>. Consulté le 26 janvier 2021

<sup>88</sup> « Fichier national d'entités », Transition bibliographique, <https://www.transition-bibliographique.fr/fne/fichier-national-entites/>. Consulté le 26 janvier 2021

<sup>89</sup> Ibid.

dans le FNE par alignement avec les référentiels d'autres métiers (éléments biographiques issus de l'Insee, numéro Siret d'entreprises faisant l'objet d'une notice "collectivités"...))<sup>90</sup>.

On le voit, par leurs missions essentielles et au-delà, les deux agences sont des acteurs cruciaux de la création de métadonnées, de données d'autorités et d'identifiants sur les productions éditoriales françaises, scientifiques et au-delà. Tout ce travail, des agences elles-mêmes et de leurs réseaux, permet une mutualisation de l'effort de catalogage et la récupération automatique par les établissements de notices bibliographiques ou d'autorités.

### *CCSD, Huma-Num, Inist, Persée : d'autres acteurs centraux de l'IST française*

Outre les agences bibliographiques, avec lesquelles les bibliothécaires travaillent le plus étroitement, d'autres acteurs de l'IST français sont à signaler, notamment le CCSD et Huma-Num (pour les sciences humaines).

Le **Centre pour la communication scientifique directe (CCSD)** administre les archives ouvertes HAL, TEL (Thèses en ligne) et Dumas (dépôt de mémoires), ainsi que les plateformes SciencesConf.org (gestion de congrès scientifiques) et Episciences.org (gestion d'épi-revues). Créé en 2000, cette structure est une unité mixte de service du CNRS. Ses missions sont notamment le développement de HAL, archive ouverte nationale, avec pour but d'interconnecter HAL aux archives institutionnelles des établissements et aux archives internationales comme arXiv ou PubMedCentral, en veillant à l'interopérabilité et à la pérennité des données<sup>91</sup>.

**Huma-Num** se définit comme « une très grande infrastructure de recherche (TGIR) visant à faciliter le tournant numérique de la recherche en sciences humaines et sociales »<sup>92</sup> qui « propose un ensemble de services pour les données numériques produites en SHS »<sup>93</sup>. Il soutient la « double démarche de mise à disposition des données de recherche (ouverture des données, des métadonnées) et d'interopérabilité des métadonnées (normalisations, API, interface d'accès aux données) »<sup>94</sup>.

Parmi ces services, on peut en citer deux particulièrement importants pour la recherche en SHS française. D'abord, un service de dépôt, de gestion et d'exposition de données, Nakala<sup>95</sup>, qui « propose deux grands types de services : des services d'accès aux données elles-mêmes et des services de présentation des métadonnées ». Ainsi, Nakala attribue des DOI aux données, permet le moissonnage des métadonnées en OAI-PMH et expose les données en RDF via un triplestore. Ensuite, le moteur de recherche Isidore<sup>96</sup>, « un service qui collecte, enrichit et offre un signalement et un accès unifié aux documents et données numériques des sciences humaines et sociales ». Il moissonne, enrichit et expose de façon ouverte « les notices, les métadonnées et le texte intégral issus

<sup>90</sup> « Le futur FNE : vers une vraie coproduction », Frédérique Joannic-Seta, *Arabesques* n° 85, 2017, <https://publications-prairial.fr/arabesques/index.php?id=246><https://publications-prairial.fr/arabesques/index.php?id=246>. Consulté le 26 janvier 2021

<sup>91</sup> « Le CCSD », CCSD, <https://www.ccsd.cnrs.fr/le-ccsd/>. Consulté le 26 janvier 2021

<sup>92</sup> « À propos de Huma-Num », Huma-Num, <https://www.huma-num.fr/>. Consulté le 26 janvier 2021

<sup>93</sup> « Services et outils », Huma-Num, <https://www.huma-num.fr/services-et-outils>. Consulté le 26 janvier 2021

<sup>94</sup> Ibid.

<sup>95</sup> « Exposer ses données avec Nakala », Huma-Num, <https://www.huma-num.fr/services-et-outils/exposer>. Consulté le 26 janvier 2021

<sup>96</sup> « Signaler ses données avec Isidore », Huma-Num, <https://www.huma-num.fr/services-et-outils/signaler>. Consulté le 26 janvier 2021



des publications électroniques, des corpus, des bases de données et des actualités scientifiques », en se fondant sur les principes du web de données.

L'**Institut de l'information scientifique et technique** (Inist) est une unité d'appui à la recherche du CNRS, dont la mission est de faciliter « l'accès, l'analyse et la fouille de l'information scientifique » et de valoriser la production scientifique<sup>97</sup>. Il a géré de 1972 à 2015 les bases bibliographiques Pascal et Francis, qui ne sont plus alimentées mais permettent toujours l'accès à plus de 20 millions de notices et quand cela est possible au texte intégral<sup>98</sup>. Aujourd'hui, il anime la plateforme Istex, base de 23 millions d'articles scientifiques acquis en licence nationale dans le cadre d'un partenariat avec le CNRS, l'Abes, le consortium Couperin et l'université de Lorraine, qui permet non seulement aux chercheurs de consulter ces documents, mais aussi d'extraire des corpus et d'exploiter le plein-texte des documents par la fouille de texte et de données<sup>99</sup>. Panist<sup>100</sup> (Plateforme d'archivage national de l'information scientifique et technique) complète l'offre d'Istex en donnant accès aux établissements éligibles aux archives d'Elsevier de 2002 à 2017 (pour le moment), soit 7,3 millions d'articles. L'Inist est également une unité support de l'équipement d'excellence Ortolang, qui veut proposer « une infrastructure en réseau offrant un réservoir de données (corpus, lexiques, dictionnaires, etc.) et d'outils sur la langue et son traitement »<sup>101</sup>.

Enfin, **Persée**, unité mixte de service dépendant du CNRS et de l'École normale supérieure de Lyon, donne accès à des corpus documentaires numérisés ou nativement numériques, et enrichis par une structuration sémantique fine. Le portail Persée présente ainsi plus de 700 000 documents, notamment en SHS, diffusés de façon ouverte.

## En Europe, une série d'initiatives articulées pour la science ouverte

L'Union européenne a lancé de nombreuses initiatives pour favoriser la science ouverte, auxquelles le PNSO propose de se rattacher, comme à EOSC, Openaire ou Go Fair.

**Horizon 2020**, devenu **Horizon Europe** début 2021, était le programme de financement de l'Union européenne pour la recherche sur la période 2014-2020, le huitième des « Programmes-cadres pour la recherche et le développement technologique »<sup>102</sup>. Les publications issues de projets financés par H2020 ont l'obligation d'être déposées en accès ouvert dans une archive ouverte ou une revue en accès ouvert<sup>103</sup>.

**Openaire** est un autre projet de la Commission européenne, né (sous le nom de Driver) en 2006, dont le but est de regrouper l'ensemble des actions de l'Union européenne en faveur de la science ouverte et notamment « de participer au développement du mandat de libre accès en Europe »<sup>104</sup>. Ses objectifs sont de « proposer

<sup>97</sup> « L'institut », Inist, <https://www.inist.fr/qui/institut/>. Consulté le 17 février 2021

<sup>98</sup> « Bases bibliographiques Pascal et Francis », Inist, <https://pascal-francis.inist.fr/>. Consulté le 17 février 2021

<sup>99</sup> « Istex », Inist, <https://www.inist.fr/services/acceder/istex/>. Consulté le 17 février 2021

<sup>100</sup> « Panist », Inist, <https://www.inist.fr/projets/panist/>. Consulté le 17 février 2021

<sup>101</sup> « Outils et Ressources pour un Traitement Optimisé de la LANGue », Ortolang, <https://www.ortolang.fr/information/presentation>. Consulté le 17 février 2021

<sup>102</sup> « Programme-cadre pour la recherche et le développement technologique », Wikipédia, [https://fr.wikipedia.org/wiki/Programme-cadre\\_pour\\_la\\_recherche\\_et\\_le\\_developpement\\_technologique](https://fr.wikipedia.org/wiki/Programme-cadre_pour_la_recherche_et_le_developpement_technologique). Consulté le 5 mars 2021

<sup>103</sup> « Le libre accès aux publications et aux données de recherche », Mesri, <https://www.horizon2020.gouv.fr/cid82025/le-libre-acces-aux-publications-aux-donnees-recherche.html>. Consulté le 5 mars 2021

<sup>104</sup> « Horizon 2020 », Wikipédia, [https://fr.wikipedia.org/wiki/Horizon\\_2020](https://fr.wikipedia.org/wiki/Horizon_2020). Consulté le 26 janvier 2021

des outils pour les dépôts des publications », « faciliter le dépôt des différents types de publication » et « travailler avec les communautés sur les différentes méthodes d'accès et de dépôt »<sup>105</sup>. Il participe également aux actions de EOSC<sup>106</sup>.

Le répertoire **Zenodo** a ainsi été créé dans le cadre d'Openaire. Il s'agit d'un dépôt généraliste de données de la recherche et de publications développé par le Cern. Zenodo revendique d'œuvrer à la science ouverte dans toutes les disciplines et d'accueillir des productions de tous les pays du monde. Les données peuvent être déposées en accès ouvert, restreint ou fermé, en revanche les métadonnées sont sous licence CC0 et peuvent être moissonnées en OAI-PMH. Zenodo pourvoit des DOI pour les ressources qui n'en disposent pas déjà, y compris à différentes versions d'une même ressource<sup>107</sup>.

**Go Fair** est une initiative prise dans le cadre de Open Science Cloud européen (EOSC), initiative de la Commission européenne pour les données de la recherche. Elle a pour objectif d'« ouvrir progressivement les données de la recherche existantes au sein des institutions scientifiques et académiques dans tous les domaines de la recherche et au-delà des frontières nationales »<sup>108</sup> et « constitue ainsi un tremplin vers la réalisation de l'Open Science Cloud européen »<sup>109</sup>. Quant à l'EOSC lui-même, il « vise à fournir une plate-forme ouverte pour l'échange de données de recherche et qui reliera les chercheurs à travers l'Europe »<sup>110</sup>.

Enfin le **Plan S**<sup>111</sup>, initiative privée mais avec des membres publics comme l'Agence nationale de la recherche, s'est développée avec surtout des acteurs européens, mais aussi au-delà. Il exige que les chercheurs bénéficiant d'un financement de ses membres publient leurs travaux en accès libre (à moyen terme) sans embargo, et définit des principes encadrant ces publications en OA, régulant les frais de publication et reconnaissant l'importance des plateformes d'auto-archivage. De nombreux éditeurs s'efforcent de se conformer aux exigences de ce plan, qui a suscité également une réaction dans les bibliothèques académiques et leurs réseaux (Sparc Europe<sup>112</sup>, ADBU<sup>113</sup>...).

## Dans le monde, un réseau d'institutions œuvrant à la structuration et à l'ouverture des métadonnées

À l'échelle mondiale, on trouve un réseau d'acteurs interconnectés entre lesquels s'opèrent des convergences, des mutualisations d'outils et de technologies de gestion des métadonnées de la recherche. La majorité des acteurs sont des organisations sans but lucratif, notamment des organismes d'attribution d'identifiants, des référentiels, des agences de normalisation... Nous présentons plusieurs exemples importants ci-dessous.

### *Crossref*

Le DOI (Digital Object Identifier) est un identifiant pérenne d'objet numérique

<sup>105</sup> Ibid.

<sup>106</sup> « Openaire in EOSC: Where we contribute », Openaire, <https://www.openaire.eu/openaire-and-eosc>. Consulté le 27 janvier 2021.

<sup>107</sup> « DOI Versioning », Zenodo, <https://help.zenodo.org/#versioning>. Consulté le 26 janvier 2021

<sup>108</sup> « Science ouverte : la France rejoint Go Fair en tant que co-fondatrice », Mesri, <https://www.enseignementsup-recherche.gouv.fr/cid124728/science-ouverte-la-france-rejoint-go-fair-en-tant-que-co-fondatrice.html>. Consulté le 26 janvier 2021

<sup>109</sup> Ibid.

<sup>110</sup> Ibid.

<sup>111</sup> « Plan S », Wikipédia, [https://fr.wikipedia.org/wiki/Plan\\_S](https://fr.wikipedia.org/wiki/Plan_S).

<sup>112</sup> « Briefing: Sparc Europe analysis of the revised Plan S », Sparc Europe, 2019, [https://sparceurope.org/briefing\\_revisedplans\\_june2019/](https://sparceurope.org/briefing_revisedplans_june2019/). Consulté le 27 janvier 2021

<sup>113</sup> « L'ADBU apporte sa contribution au Plan S ! », ADBU, 2019, <https://adbu.fr/adbu-apporte-sa-contribution-au-plan-s/>. Consulté le 27 janvier 2021

normalisé depuis 2012 par la norme ISO 26324:2012. Environ 238 millions de DOI avaient été attribués au 18 janvier 2021<sup>114</sup>, toutes agences confondues. Crossref est la principale de ces agences. L'organisation, créée en 2000, revendique 15 700 membres (éditeurs principalement, mais aussi bibliothèques, archives institutionnelles, équipes de recherche...), et a déjà attribué des identifiants à près de 120 millions de documents (en janvier 2021)<sup>115</sup>.

En attribuant un DOI à une ressource (article, acte de conférence, prépublication...), Crossref récupère un ensemble de métadonnées (le titre, le nom des auteurs et le nom de la revue...), qu'elle expose ensuite de façon ouverte. En effet, les éditeurs ou autres pourvoyeurs de données renseignent de multiples métadonnées, bibliographiques et au-delà (financements, licence, lien vers le full-text, etc.). De plus en plus, les références citées par les articles sont également déposées par les éditeurs et exposées (dans le cadre d'I4OC)<sup>116</sup>. Crossref n'accomplit pas de corrections sur les métadonnées envoyées par les éditeurs, mais peut en revanche les enrichir, par exemple en établissant des liens manquants avec des ressources citées ou en ajoutant des financeurs. Crossref propose aux tiers plusieurs API pour récupérer ces métadonnées. L'archive nationale HAL fait par exemple usage de ce service pour récupérer automatiquement des métadonnées, ce qui permet de faciliter le dépôt d'un article au moyen de son DOI<sup>117</sup>.

Crossref est une association sans but lucratif dont le financement repose essentiellement sur les cotisations des membres et les frais d'enregistrement de contenus, ainsi que sur la facturation de services payants additionnels (environ 9 millions de dollars US de revenus récoltés en 2019)<sup>118</sup>. Les membres fondateurs sont cependant notamment de grands éditeurs lucratifs<sup>119</sup>.

Crossref est ainsi un nœud crucial au sein d'un réseau complexe où s'échangent les métadonnées, permettant de faire des liens entre les entités (ressources, acteurs, organisations...), de récolter les métadonnées à grande échelle auprès des sources (éditeurs...), pour les mettre à disposition d'un ensemble de bases de données. Il rend ainsi possible une meilleure exposition de ces métadonnées, et donc des données elles-mêmes<sup>120</sup>. Des obstacles se posent cependant pour que Crossref expose les métadonnées des publications scientifiques de façon plus large encore : nombre de publications n'ont pas de DOI (ou pas attribués par Crossref), et si c'est le cas, de nombreuses métadonnées peuvent manquer, comme les résumés, les affiliations des auteurs (cruciales pour les indicateurs des établissements)<sup>121</sup>.

---

<sup>114</sup> « Frequently Asked Questions about the DOI® System », DOI.org, <https://www.doi.org/faq.html>. Consulté le 17 février 2021

<sup>115</sup> « New public data file: 120+ million metadata records », Crossref, 2021, <https://www.crossref.org/blog/new-public-data-file-120-million-metadata-records/>. Consulté le 17 février 2021

<sup>116</sup> « Open Metadata of Scholarly Publications », Ludo Waltman, 2019, [https://ec.europa.eu/info/sites/info/files/research\\_and\\_innovation/open\\_metadata\\_of\\_scholarly\\_publications\\_0.pdf](https://ec.europa.eu/info/sites/info/files/research_and_innovation/open_metadata_of_scholarly_publications_0.pdf). Consulté le 26 janvier 2021

<sup>117</sup> « Déposer », HAL, <https://doc.archives-ouvertes.fr/deposer/>. Consulté le 26 janvier 2021

<sup>118</sup> « Annual report 2019 », Crossref, <https://www.crossref.org/annual-report/#2019>. Consulté le 26 janvier 2021

<sup>119</sup> Elsevier Science, Academic Press, Inc., IEEE, Springer Verlag, Kluwer Academic Publishers, Nature Publishing Group, Oxford University Press, John Wiley & Sons, Inc., Blackwell Science, ainsi que American Association for the Advancement of Science, American Institute of Physics, Association for Computing Machinery

<sup>120</sup> Le Plan national pour la science ouverte promeut une plus grande participation dans le développement de Crossref. Voir « Des identifiants ouverts pour la science ouverte : note d'orientation », Comité pour la science ouverte, 2019, <https://www.ouvrirlascience.fr/des-identifiants-ouverts-pour-la-science-ouverte-note-orientation/>. Consulté le 26 janvier 2021. Par ailleurs, notons que Marin Dacos est membre du board de Crossref, pour OpenEdition.

<sup>121</sup> Ibid.

### *Datacite*

Datacite est l'un des principaux pourvoyeurs mondiaux de DOI, quoique de façon plus modeste que Crossref : il a à ce jour identifié plus de 21 millions de travaux<sup>122</sup>. Il s'agit d'une organisation sans but lucratif, dont les membres viennent avant tout du monde académique<sup>123</sup>, à la différence de Crossref. Datacite identifie avant tout des jeux de données de la recherche, tandis que Crossref identifie plutôt des publications (articles, livres...). À ce titre, les deux organisations revendiquent d'être complémentaires plus que concurrentes, et collaborent sur de nombreux projets (Freya, ROR, PIDapalooza, Conference identifiers, Metadata 2020). Par ailleurs, leurs modèles économiques et politiques tarifaires différentes leur impose de coopérer pour coexister.

Datacite propose également un registre international des entrepôts de données de la recherche (Re3data<sup>124</sup>) permettant d'identifier selon les disciplines et les institutions l'entrepôt de données le plus approprié.

Notons que H2020 inclue Datacite dans ses recommandations pour l'accès ouvert : « *The Datacite service is recommended in the "European Commission's Guidelines on Open Access to Scientific Publications and Research Data in Horizon 2020", which states that when providing open access to publications in repositories "where possible, contributors should also be uniquely identifiable, and data uniquely attributable, through identifiers which are persistent, non-proprietary, open and interoperable (e.g. through leveraging existing sustainable initiatives such as Orcid for contributor identifiers and Datacite for data identifiers).* »<sup>125</sup>

### *Le Centre international d'enregistrement des publications en série (Cieps) et l'ISSN*

Le Cieps, centre international en charge de l'ISSN (numéro international normalisé des publications en série) créé en 1976, est « une organisation intergouvernementale ayant pour fonction de coordonner au niveau international l'identification et la description des publications en série et ressources continues, imprimées et en ligne, dans toutes les disciplines »<sup>126</sup>.

Dépendant de l'Unesco, il est hébergé par la France et largement financé par ces deux acteurs et les États membres<sup>127</sup>, ainsi que par des ressources propres apportées via des prestations de services aux éditeurs et aux bibliothèques.

L'ISSN est normalisé par la norme ISO 3297:2017<sup>128</sup>. Selon le rapport d'activité

<sup>122</sup> « Statistics », DataCite Commons, <https://commons.datacite.org/statistics>. Consulté le 18 février 2021

<sup>123</sup> Les membres de DataCite à sa création le 1<sup>er</sup> décembre 2009 en témoignent : British Library (Royaume-Uni), Technical Information Center of Denmark (Danemark), TU Delft Library (Pays-Bas), National Research Council's Canada Institute for Scientific and Technical Information (Canada), California Digital Library (États-Unis), Purdue University (États-Unis), German National Library of Science and Technology (Allemagne), rejoints rapidement ensuite par ETH Zurich Library (Suisse), Institut de l'information scientifique et technique (France), Australian National Data Service (Australie), GESIS - Leibniz Institute of Social Sciences (Allemagne), Conseil national de recherches Canada (Canada). Voir « DataCite », Wikipédia, <https://en.wikipedia.org/wiki/DataCite>. Consulté le 26 janvier 2021

<sup>124</sup> Re3Data, <https://www.re3data.org/>. Consulté le 26 janvier 2021

<sup>125</sup> « Open access briefing paper: The potential of global identifiers to support more efficient workflows for all kinds of OA », Jisc, 2018, <https://scholarlycommunications.jiscinvolve.org/wp/2018/10/24/open-access-briefing-paper-the-potential-of-global-identifiers-to-support-more-efficient-workflows-for-all-kinds-of-oa/>. Consulté le 26 janvier 2021

<sup>126</sup> « Le Centre international d'enregistrement des publications en série », ISSN, <https://www.issn.org/fr/le-centre-et-le-reseau/notre-mission/presentation-du-centre/>. Consulté le 26 janvier 2021

<sup>127</sup> « Liste des pays membres 2019 », ISSN, [https://www.issn.org/wp-content/uploads/2019/10/LISTE-DES-PAYS-MEMBRES\\_2019.pdf](https://www.issn.org/wp-content/uploads/2019/10/LISTE-DES-PAYS-MEMBRES_2019.pdf). Consulté le 26 janvier 2021

<sup>128</sup> « ISO 3297:2017 Information et documentation — Numéro international normalisé des publications en série (ISSN) », Organisation internationale de normalisation, <https://www.iso.org/fr/standard/73322.html>. Consulté le 26 janvier 2021

2019 du Cieps, le registre international de l'ISSN contenait 2 121 389 notices fin 2019 et 58 275 nouvelles notices y avaient été ajoutées en 2019. Le nombre d'attributions d'ISSN aux ressources numériques est en croissance, avec 22 705 nouvelles ressources en ligne identifiées en 2019, soit 39 % du total d'ISSN attribués cette année-là. Un total de 269 868 ressources en ligne étaient recensées dans le registre ISSN en janvier 2020<sup>129</sup>.

Le Cieps développe le service Road depuis fin 2013, pour donner « accès gratuitement à un sous-ensemble de notices bibliographiques du Portail ISSN qui décrivent des ressources scientifiques en libre accès identifiées par un ISSN »<sup>130</sup>, notices qui sont « enrichies par des métadonnées issues de bases d'indexation, de répertoires (DOAJ, Latindex, The Keepers Registry) et d'indicateurs de performance (Scopus) »<sup>131</sup>.

### Orcid

L'identifiant Orcid (Open Researcher and Contributor ID) est le principal identifiant pérenne et ouvert pour les chercheurs et contributeurs de publications scientifiques à l'échelle internationale. Il s'agit d'une URI à 16 chiffres compatible avec le standard ISO 27729:2012, soit l'identifiant international normalisé du nom (Isni).

L'identifiant Orcid fonctionne par auto-enregistrement (gratuit) des contributeurs. Malgré la puissance de son caractère international et son partage déjà large, il faut donc souligner les problèmes de qualité qui peuvent être occasionnés par l'absence de curation professionnelle systématique. Il est potentiellement associé à différentes informations (institutions, projets, publications...) que le chercheur peut choisir de partager publiquement ou pas. Il constitue un élément crucial de l'identité numérique des chercheurs dans de nombreuses disciplines, mais encore trop peu dans d'autres, notamment en SHS.

Interopérable, il permet à des systèmes tiers de recueillir des informations dans Orcid, ou inversement. Il est par exemple possible pour un auteur de verser ses références de dépôts HAL dans Orcid. Les bases de données peuvent aussi y recueillir des informations, ce qui permet parfois de pallier des manques dans les métadonnées des éditeurs<sup>132</sup>.

Orcid est une organisation sans but lucratif inscrite dans le Delaware, financée par les cotisations de ses membres ainsi que par des services payants. Les membres sont au nombre de 1 013 selon le rapport 2019<sup>133</sup>. En février 2021 Orcid dénombrait 10 786 808 identifiants distribués<sup>134</sup>. En 2019, selon le rapport annuel, 1 957 249 nouveaux enregistrements avaient eu lieu, soit environ 5400 par jour, une augmentation de 8 % par rapport à 2018<sup>135</sup>. Cette croissance importante s'explique notamment par l'exigence d'un identifiant Orcid de la part de certains financeurs et éditeurs. Ces pratiques permettent de contrebalancer progressivement un constat fait 2017 : « *An analysis undertaken by Jisc of Crossref metadata found that of the 14.6 million non-unique authors of journal articles*

<sup>129</sup> « Rapport d'activité du centre international de l'ISSN pour l'année 2019 », Cieps, 2020, [https://www.issn.org/wp-content/uploads/2020/06/ISSN\\_RAPPORT\\_ACTIVITE\\_2019\\_FR\\_FINAL\\_10Juin.pdf](https://www.issn.org/wp-content/uploads/2020/06/ISSN_RAPPORT_ACTIVITE_2019_FR_FINAL_10Juin.pdf). Consulté le 18 février 2021

<sup>130</sup> « Road, le répertoire des ressources scientifiques et universitaires en libre accès », ISSN, <https://www.issn.org/fr/services-et-prestations/services-en-ligne/road/>. Consulté le 26 janvier 2021

<sup>131</sup> Ibid.

<sup>132</sup> « Availability of Orcids in Publications Archived in PubMed, Medline, and Web of Science Core Collection », Christophe Boudry, *Scientometrics*, 2021, <https://link.springer.com/article/10.1007/s11192-020-03825-7>. Consulté le 18 février 2021

<sup>133</sup> « Orcid 2019 Annual Report », Orcid / Laurel, Haak; Petro, Julie Anne; Simpson, Will; Demeranville, Tom; Wijnbergen, Ivo; Hershberger, Sarah; et al., 2020, <https://doi.org/10.23640/07243.12009153.v1>. Consulté le 18 février 2021

<sup>134</sup> « Statistics », DataCite Commons, <https://commons.datacite.org/statistics>. Consulté le 18 février 2021

<sup>135</sup> « Orcid 2019 Annual Report », Op. Cit.

*published in 2017, less than 900,000 (6%) had an Orcid associated with their record.* »<sup>136</sup> Cela s'expliquait en partie par l'absence d'identifiant chez certains chercheurs, mais aussi par l'incapacité de certains éditeurs à transmettre les identifiants Orcid correctement à Crossref. L'enjeu ne se limite donc pas à celui de l'attribution d'un identifiant aux chercheurs, mais est aussi celui du partage effectif de ces identifiants grâce à des protocoles standardisés entre les acteurs et l'interopérabilité des systèmes.

En France, conformément au vœu du PNSO, une « communauté française Orcid France est née fin 2019 et comprend actuellement 36 membres »<sup>137</sup>. Le développement de l'Orcid dans la recherche française est donc considéré comme crucial, avec une adhésion nationale, mais aussi pensée en interconnexion avec d'autres identifiants nationaux, tels l'IdRef ou l'idHAL.

### *Viaf*

Le fichier d'autorité international virtuel Viaf (Virtual International Authority File), administré par OCLC, sert à identifier, essentiellement, des personnes et des collectivités. Il met en relation les notices d'autorité des différentes bibliothèques concernant une même personne. Il est alimenté en fichiers d'autorité par plus de 50 organisations, dont des bibliothèques nationales, et « établit ensuite des correspondances et fusionne les notices en de “super” notices d'autorité »<sup>138</sup>.

Lancé en 2003 par l'OCLC, la Library of Congress et la Bibliothèque nationale allemande, le Viaf a été rejoint en 2009 par la BnF. Le Viaf est également lié aux données de Wikipédia et Wikidata.

### *Isni*

L'Isni (International Standard Name Identifier) est la norme certifiée ISO 27729:2012 permettant « d'identifier au niveau international les identités publiques des personnes ou des organismes impliqués dans la création, la production ou la gestion et la distribution de contenus intellectuels et artistiques »<sup>139</sup>. Elle est accessible gratuitement depuis 2011<sup>140</sup>.

L'Isni est géré par une agence internationale d'attribution et par un réseau d'agences d'enregistrement qui servent d'intermédiaires avec les déclarants, à l'échelon national. La BnF et l'Abes font partie de ce réseau depuis janvier 2014<sup>141</sup>.

---

<sup>136</sup> « The potential of global identifiers to support more efficient workflows for all kinds of OA », Chris Brown, Katie Shamash, Helen Blanchett, Balviar Notay, 2018, <http://repository.jisc.ac.uk/7089/1/2018JiscOABriefingPotentialGlobalIdentifiers.pdf>. Consulté le 27 janvier 2021

<sup>137</sup> « Consortium Orcid France : nouveau réseau, nouveaux correspondants », Isabelle Mauger Perez, *Arabesques* n° 97, 2020, <https://publications-prairial.fr/arabesques/index.php?id=1772>. Consulté le 27 janvier 2021

<sup>138</sup> « Fichiers d'autorités pour simplifier le catalogage », OCLC, <https://www.oclc.org/fr/worldcat/cooperative-quality.html>. Consulté le 27 janvier 2021

<sup>139</sup> « Isni (International Standard Name Identifier) », BnF, <https://www.bnf.fr/fr/isni-international-standard-name-identifieur>. Consulté le 27 janvier 2021

<sup>140</sup> « Search Database », Isni, <https://isni.org/page/search-database/>. Consulté le 27 janvier 2021

<sup>141</sup> Ibid.

### *ROR et Ringgold*

ROR (Research Organization Registry)<sup>142</sup> a été lancé en janvier 2019 pour pallier les faiblesses des identifiants de structures associées à la recherche. Ses principaux instigateurs et responsables actuels sont la California Digital Library, Crossref, Datacite et Digital Science. Le référentiel est fondé sur les données de Grid (Global Research Identifier Database)<sup>143</sup>, piloté par Digital Science, et référence près de 100 000 institutions de recherche<sup>144</sup>.

Ses données sont diffusées en CC0, mais des services payants devraient être mis en place à partir de 2022<sup>145</sup>, sans remettre en cause la gratuité de ces données. L'identifiant ROR est associé à un ensemble de métadonnées rapportant des informations sur l'organisation identifiée. Il est aligné sur d'autres identifiants : Grid, Isni, Crossref Funder ID et Wikidata.

Ringgold<sup>146</sup> est un identifiant utilisé par les éditeurs pour identifier les acteurs de la chaîne de l'édition et faciliter les échanges entre les acteurs. Il s'agit en conséquence de la plus grosse base mondiale de structures universitaires et de recherche. Si les bibliothécaires n'ont pas de rapport direct avec ce système d'identifiants, il convient d'être attentif à d'éventuels alignements avec les bases Orcid, ROR, ou autres.

### *OCLC*

OCLC, Inc., fondé en 1967 sous le nom de Ohio College Library Center, puis renommé Online Computer Library Center, est une organisation sans but lucratif basée aux États-Unis, coopérative de services du monde des bibliothèques à l'échelle internationale. Des observateurs voient cependant en elle « *la posture d'une multinationale* »<sup>147</sup>. Son financement est assuré par la facturation aux bibliothèques clientes de ses multiples services. L'organisation revendique 15 637 membres dans 107 pays en février 2021<sup>148</sup>.

OCLC développe notamment le catalogue mondial de bibliothèques Worldcat. De plus, OCLC est en charge du fichier Vial susmentionné, il est à l'origine du vocabulaire de métadonnées Dublin Core (et l'a géré jusqu'en 2008) et administre depuis 1988 la classification décimale Dewey. Enfin, OCLC propose également un outil de découverte, Worldcat Discovery, ou encore un logiciel de proxy, EZproxy, utilisé en bibliothèque universitaire, parmi une gamme de services (payants) assez étendue.

Le catalogue mondial Worldcat (lancé en 1971) rassemblait en janvier 2021 quelque 503 millions de notices bibliographiques pointant sur 3 milliards de documents<sup>149</sup>, ce qui en fait la plus grande base de données bibliographique du monde. Worldcat rassemble des métadonnées apportées par les bibliothèques membres, qui permettent aux utilisateurs de trouver des documents depuis le site web Worldcat.org, et d'identifier les bibliothèques

---

<sup>142</sup> « About ROR », ROR, <https://ror.org/about/>. Consulté le 27 janvier 2021

<sup>143</sup> « Grid – Global Research Identifier Database », Grid, <https://grid.ac/>. Consulté le 27 janvier 2021

<sup>144</sup> « What Does it Mean to Be in ROR? », Maria Gould / ROR, 2020, <https://ror.org/blog/2020-11-18-what-does-it-mean-to-be-in-ror/>. Consulté le 27 janvier 2021

<sup>145</sup> « Governance », ROR, <https://ror.org/governance/>. Consulté le 27 janvier 2021

<sup>146</sup> « Ringgold ID (P3500) », Wikidata, <https://www.wikidata.org/wiki/Property:P3500>. Consulté le 25 février 2021

<sup>147</sup> Philippe Bourdenet, « OCLC, l'histoire d'une coopération fructueuse », in *Documentaliste-Sciences de l'information*, vol. 50, n° 2, juillet 2013, p. 30.

<sup>148</sup> « About », OCLC, <https://www.oclc.org/fr/about.html>. Consulté le 20 février 2021

<sup>149</sup> « Inside WorldCat », OCLC, <https://www.oclc.org/en/worldcat/inside-worldcat.html>. Consulté le 20 février 2021

où ces documents sont disponibles dans un espace géographique donné.

La BnF s'est rattachée à Worldcat en 2009 (et a alors versé 13 millions de notices de son catalogue général), comme l'Abes (qui a alors versé 9 millions de notices bibliographiques Sudoc). Ainsi, quelque 35 millions de documents de bibliothèques françaises seraient référencés dans Worldcat (71 millions d'exemplaires), ce qui est important mais sous-tendu par des problèmes de doublonnage ou de mise à jour (le Catalogue collectif de France décompte de son côté 30 millions de documents<sup>150</sup>). Néanmoins, selon David Aymonin<sup>151</sup>, directeur de l'Abes, des obstacles, comme des problèmes d'échanges de données ou de normes, ainsi que de frais d'adhésion, bloquent la participation d'un certain nombre de bibliothèques françaises, qui ne seraient que 2 000 dans Worldcat, tandis que le Catalogue collectif de France en dénombre 5 000. Un travail d'analyse entre l'OCLC, la BnF et l'Abes est donc nécessaire pour améliorer le transfert de données, pour pouvoir développer de nouveaux services, notamment de prêt international entre bibliothèques, qui permettrait de pallier partiellement le déclin des budgets d'acquisition partout dans le monde.

### *Niso*

Niso (National Information Standards Organization) est un organisme de normalisation américain dans le domaine de l'information. Il s'agit d'un organisme sans but lucratif accrédité par l'American National Standards Institute (Ansi). Il a vocation à « initier, développer, maintenir et publier des normes techniques pour les services d'information, les bibliothèques, les éditeurs et autres acteurs impliqués dans des activités de création, stockage, conservation, partage, accès et diffusion d'information, et ceci quel que soit le type de média (texte, image, son, audiovisuel...) ou le vecteur (numérique ou physique) »<sup>152</sup>.

Quoique l'organisme soit américain, de nombreuses normes qu'il a édictées ont un impact international. C'est le cas par exemple de la norme Z39.50 permettant l'interrogation de bases de données, initiée par Niso avant d'être maintenue par la Library of Congress. Le vocabulaire Dublin Core est défini par la norme Ansi/Niso Z39.85<sup>153</sup>. Niso a aussi normalisé Jats<sup>154</sup>, Kbart<sup>155</sup>, ou encore la syntaxe des DOI<sup>156</sup>.

## **Un écosystème international dynamique, dans lequel la France est encore peu impliquée**

Des initiatives multiples émergent au niveau mondial, notamment autour des institutions précédemment mentionnées, dont Crossref. De façon générale, on peut

---

<sup>150</sup> « Les bibliothèques françaises dans Worldcat : l'Abes souhaite une concertation avec la BnF et OCLC », Archimag, 2019, <https://www.archimag.com/bibliotheque-edition/2019/03/19/bibliotheques-fran%C3%A7aises-worldcat-abes-concertation-bnf-oclc>. Consulté le 5 mars 2021

<sup>151</sup> Ibid.

<sup>152</sup> « National Information Standards Organization », Wikipédia, [https://fr.wikipedia.org/wiki/National\\_Information\\_Standards\\_Organization#cite\\_note-2](https://fr.wikipedia.org/wiki/National_Information_Standards_Organization#cite_note-2). Consulté le 29 janvier 2021

<sup>153</sup> « Ansi/Niso Z39.85-2012 The Dublin Core Metadata Element Set », Niso, <https://www.niso.org/publications/ansiniso-z3985-2012-dublin-core-metadata-element-set>. Consulté le 29 janvier 2021

<sup>154</sup> « Ansi/Niso Z39.96-2019, Jats: Journal Article Tag Suite, version 1.2 », Niso, <https://www.niso.org/publications/z3996-2019-jats>. Consulté le 29 janvier 2021

<sup>155</sup> « Niso RP-9-2014, Kbart: Knowledge Bases and Related Tools Recommended Practice », Niso, <https://www.niso.org/publications/rp-9-2014-kbart>. Consulté le 29 janvier 2021

<sup>156</sup> « Niso Z39.84-2005 (R2010) - Syntax for the Digital Object Identifier », Techstreet, [https://www.techstreet.com/standards/niso-z39-84-2005-r2010?product\\_id=1262088](https://www.techstreet.com/standards/niso-z39-84-2005-r2010?product_id=1262088). Consulté le 29 janvier 2021



souligner la présence peu visible encore des institutions françaises dans ces initiatives. Néanmoins, le collège Europe et international du Comité pour la science ouverte a lancé le 1<sup>er</sup> février 2021 un appel à manifestation d'intérêt pour la construction d'un Réseau d'experts internationaux de la science ouverte (Reiso), car afin de « mener à bien une politique de science ouverte dans un contexte européen et international, il est nécessaire que la France renforce ou installe sa représentativité dans les instances, organisations et événements significatifs au niveau international »<sup>157</sup>. Il s'agit d'assurer à terme une présence française dans les *boards* des organisations internationales de la science ouverte, et une représentation dans les rencontres internationales pour y faire valoir le positionnement français. Il est trop tôt pour évaluer l'impact concret de cette initiative, qui témoigne cependant d'un volontarisme au niveau national.

Parmi les initiatives internationales récentes, **Principles for Open Scholarly Infrastructure**<sup>158</sup> a établi en 2020 une liste de principes en termes de gouvernance, de transparence, d'ouverture des données et de pérennité du modèle économique. L'initiative a été rejointe par Crossref en novembre 2020<sup>159</sup>.

Un autre exemple est **Metadata2020**<sup>160</sup> : cette initiative réunit des institutions diverses (universités, dépôts de données, agences d'identifiants, éditeurs privés, bibliothèques...) pour soutenir des initiatives en faveur de métadonnées de meilleure qualité et plus ouvertes. Elle a produit un certain nombre de ressources en ce sens, avec une structuration selon des « communautés » (chercheurs, éditeurs, bibliothécaires, etc.)<sup>161</sup>.

Un autre exemple est l'événement **PIDapalooza**<sup>162</sup>, organisé chaque année depuis 2016 (sauf en 2017)<sup>163</sup> notamment par la California Digital Library, Crossref, Datacite, Orcid et Niso. Un ensemble de conférences autour des identifiants pérennes et des métadonnées se tient en présentiel ou distanciel (en 2021), et permet des retours d'expériences et des échanges de bonnes pratiques.

Des initiatives à l'impact concret déjà perceptibles sont OpenCitations et son initiative I4OC, et le projet cousin I4OA. Elles ont pour objectifs de rendre plus ouvertes les données de citations des articles scientifiques, et leurs résumés, deux ensembles de métadonnées cruciaux pour la recherche.

**OpenCitations** est une initiative du Jisc lancée en 2010, mais aujourd'hui gérée par le Research Centre for Open Scholarly Metadata de l'université de Bologne. Son projet Coci<sup>164</sup>, l'OpenCitations Index of Crossref open DOI-to-DOI citations, lancé le 4 juillet 2018, est, comme son nom l'indique, un index de citations issus des données renseignées dans Crossref. Le modèle s'appuie sur l'ontologie CITO<sup>165</sup> qui fait partie des ontologies Spar (créées par OpenCitations).

<sup>157</sup> « AMI pour la construction d'un réseau d'experts internationaux de la Science ouverte (ReiSo) », Ouvrir la science, 2021, <https://www.ouvrirlascience.fr/ami-reseau-experts-internationaux-de-la-science-ouverte/>. Consulté le 3 février 2021

<sup>158</sup> « The Principles of Open Scholarly Infrastructure », Bilder G., Lin J., Neylon C., <https://doi.org/10.24343/C34W2H>. Consulté le 27 janvier 2021

<sup>159</sup> « Crossref's Board votes to adopt the Principles of Open Scholarly Infrastructure », Geoffrey Bilder / Crossref, 2020, <https://www.crossref.org/blog/crossrefs-board-votes-to-adopt-the-principles-of-open-scholarly-infrastructure/>. Consulté le 27 janvier 2021.

<sup>160</sup> « About », Metadata 2020, <http://www.metadata2020.org/about/>

<sup>161</sup> « Communities », Metadata 2020, <http://www.metadata2020.org/communities/>

<sup>162</sup> « The PIDapalooza Community », PIDapalooza, <https://www.pidapalooza.org/about>

<sup>163</sup> « Explore PIDapaloozas of Years Past », PIDapalooza, <https://www.pidapalooza.org/past-events>

<sup>164</sup> « Software review: COCI, the OpenCitations Index of Crossref open DOI-to-DOI citations », Heibi, I., Peroni, S. & Shotton, D., *Scientometrics* n° 121, 2019, <https://doi.org/10.1007/s11192-019-03217-6>. Consulté le 29 janvier 2021

<sup>165</sup> « Citation Typing Ontology (Cito) », Spar, <http://www.sparontologies.net/ontologies/cito>. Consulté le 29 janvier 2021

Le but est de créer un index ouvert des citations des articles scientifiques afin de proposer une alternative aux données de citation des bases de données privées (notamment l'Impact Factor de Clarivate-Web of Science), qui ont un impact important notamment sur l'évaluation de la recherche. L'obstacle principal à cette initiative est l'absence d'accès ouvert aux références d'un certain nombre d'articles sur Crossref, parce que les éditeurs ne renseignent pas ces références ou refusent de les rendre ouvertes.

C'est pourquoi l'initiative **I4OC** (Initiative for Open Citations<sup>166</sup>), lancée en 2017 par OpenCitations et d'autres organisations (Wikimedia Foundation, Plos, eLife, Datacite, et le Centre for Culture and Technology de la Curtin University) œuvre à ce que les éditeurs rendent ouvertes leurs données de citations afin que l'initiative Coci puissent référencer un plus grand nombre d'entre elles.

L'initiative a eu du succès, le pourcentage de citations ouvertes dans Crossref est passé, selon I4OC, de 1 % en 2016 à 84 % en janvier 2021<sup>167</sup>, grâce à un activisme auprès d'éditeurs qui refusaient encore de rendre les listes de références de leurs articles ouvertes dans Crossref. Même Elsevier, qui s'était d'abord montré réticent à se joindre l'initiative, a annoncé l'ouverture de ses données de citations dans Crossref en décembre 2020<sup>168</sup>. Le Plan national pour la science ouverte<sup>169</sup> et le Plan S<sup>170</sup> soutiennent cette initiative.

Une autre initiative récente, lancée en septembre 2020, est l'**Initiative for Open Abstracts (I4OA)**<sup>171</sup>, qui veut promouvoir la mise à disposition ouverte par les éditeurs via Crossref des résumés des articles scientifiques, afin qu'ils soient non seulement accessibles pour les chercheurs (ce qui peut déjà être le cas sur le site des éditeurs), mais aussi qu'ils puissent être exploités automatiquement par des machines, à large échelle, à des fins de revues systématiques ou de fouille de texte, ou tout simplement pour améliorer l'efficacité des outils de découverte<sup>172</sup>. Cela nécessite l'ouverture légale de ces données, mais aussi leur mise à disposition réelle pour les machines, ce qui suppose une infrastructure technique adaptée, que peut apporter Crossref.

Or les résumés sont souvent manquants dans Crossref. Ainsi, sur cette plateforme, seules 8 % des publications ont un résumé disponible (20 % sur la période 2018-2020 cependant<sup>173</sup>), un chiffre bien plus faible que sur le Web of Science ou Scopus<sup>174</sup>. L'enregistrement par les éditeurs des résumés de leurs publications dans Crossref permettrait l'accès à ces résumés à large échelle via une API unique, sous un format uniforme lisible par les machines, et non éparpillés à travers une multitude de sites d'éditeurs, ou de bases de données, sous des formats divers. L'initiative I4OA, plus récente qu'I4OC, revendiquait fin 2020 l'adhésion de 40 éditeurs, d'importances variables.

---

<sup>166</sup> « Initiative for Open Citations », I4OC, <https://i4oc.org/>. Consulté le 29 janvier 2021

<sup>167</sup> Tweet d'I4OC du 20 janvier 2021, [https://twitter.com/i4oc\\_org/status/1351922437584269316?s=20](https://twitter.com/i4oc_org/status/1351922437584269316?s=20). Consulté le 29 janvier 2021

<sup>168</sup> « Advancing Responsible Research Assessment », Elsevier / Andrew Plume, 2020, <https://www.elsevier.com/connect/advancing-responsible-research-assessment>. Consulté le 21 février 2021

<sup>169</sup> « Plan national pour la science ouverte », Comité pour la science ouverte, 2019, <https://www.ouvrirelascience.fr/plan-national-pour-la-science-ouverte>. Consulté le 29 janvier 2021

<sup>170</sup> « Principles and Implementation », Plan S, <https://www.coalition-s.org/addendum-to-the-coalition-s-guidance-on-the-implementation-of-plan-s/principles-and-implementation/>. Consulté le 29 janvier 2021

<sup>171</sup> « Initiative for Open Abstracts Launches to Promote Discovery of Research », I4OA, <https://i4oa.org/press.html#pressrelease>. Consulté le 29 janvier 2021

<sup>172</sup> « Why openly available abstracts are important – Overview of the current state of affairs », Aaron Tay, 2020, <https://medium.com/a-academic-librarians-thoughts-on-open-access/why-openly-available-abstracts-are-important-overview-of-the-current-state-of-affairs-bb7bde1ed751>. Consulté le 29 janvier 2021

<sup>173</sup> Ibid.

<sup>174</sup> « Open Metadata of Scholarly Publications », Ludo Waltman, 2019, [https://ec.europa.eu/info/sites/info/files/research\\_and\\_innovation/open\\_metadata\\_of\\_scholarly\\_publications\\_0.pdf](https://ec.europa.eu/info/sites/info/files/research_and_innovation/open_metadata_of_scholarly_publications_0.pdf). Consulté le 29 janvier 2021

## Vers une base de données ouverte et mondiale des publications scientifiques ?

Nous le voyons, on dispose aujourd'hui d'un ensemble d'institutions émettant des standards, des bonnes pratiques, des identifiants pérennes largement partagés à l'échelle mondiale, et d'initiatives collectives pour ouvrir toujours plus les données et les métadonnées dans le cadre du web sémantique. Le travail des professionnels de l'IST, aux échelles de l'établissement pour les bibliothèques universitaires et du pays pour les agences bibliographiques, est de créer une dynamique pour que les chercheurs et les institutions, au premier rang desquelles les bibliothèques, s'approprient ces standards, bonnes pratiques et identifiants. Ce socle est indispensable pour aller vers un objectif de plus en plus tangible à l'échelle mondiale, celle d'un index libre et général des données et métadonnées des publications scientifiques, facilitant le repérage univoque et l'accès universel aux publications, mais également susceptible d'être une alternative aux grandes bases propriétaires et très coûteuses que sont Scopus ou Web of Science, ou à Google Scholar. Le cadre du web sémantique favorise un tel objectif, dont on pourrait voir les prémices dans une base de connaissances comme Wikidata. Pour y voir plus clair, nous présentons dans les prochaines pages les diverses bases de données dans lesquelles les métadonnées des productions scientifiques sont recueillies et exposées.

### *La grande aventure des bases de données bibliographiques*

Depuis les années 1960, puis plus encore à partir du milieu des années 1970, l'informatique en général et l'informatique documentaire en particulier ont pris leur essor, permettant l'interrogation de « gisements documentaires » à distance, via des réseaux téléphoniques ou des lignes spécialisées. La brochure « Des banques de données pour les étudiants, les enseignants et les chercheurs »<sup>175</sup>, dont nous avons consulté l'édition de 1994, montre le bouleversement dans les pratiques de recherche documentaire amené par ces nouvelles technologies, et les opportunités pour la constitution de bibliographie, la veille disciplinaire, et l'accès à des textes intégraux et des données.

Préalablement au développement de l'informatique, qui a permis la multiplication de ces bases, de grands répertoires bibliographiques, tels Chemical Abstracts, fondé en 1907, ou Medline, née en 1946, avaient permis d'industrialiser la production d'outils d'information secondaire, diffusés mondialement. L'apport des outils informatisés a constitué une vraie révolution des méthodes de bibliographie pour les étudiants et les chercheurs, et bien sûr pour les producteurs de banques de données. Un décompte des banques de données disponibles à travers le monde indique un passage de 300 en 1979 à 5 210 en 1993.

Les coûts de production des bases générés par le travail de collecte des publications papier, leur analyse et la création des données informatisées (saisie, indexation, lexiques, thésaurus, identifiants de produits et substances, formules chimiques et mathématiques, harmonisation et codage des unités de mesure, réécriture et normalisation des résumés, etc.), et la mise à disposition des outils techniques (serveurs, langages d'interrogation

---

<sup>175</sup> « Des banques de données pour les étudiants, les enseignants et les chercheurs », Clarisse Marandin, MESR, 1994, 6<sup>e</sup> édition mise à jour.

– booléens, opérateurs de proximité... –, gestion des abonnements et des consommations, diffusion des supports, formation, documentation, etc.) justifiaient des tarifs élevés et exigeaient une maîtrise de l'interrogation qui s'est progressivement démocratisée grâce au support CD-Rom.

Une liste des banques de données disponibles en France en 1994 montre la spécialisation extrêmement poussée de certaines de ces bases : aux côtés de Francis (SHS) et Pascal (STM) de l'Inist, de Medline (médecine) de la National Library of Medicine (États-Unis), ou de CAS Chemical Abstracts (chimie) de l'American Chemical Society, on trouve par exemple Adhemix, du CEA, spécialisée dans les « colles du marché français ». La plupart des grandes bases sont encore consultables par Internet, Chem Abs devenue SciFinder, Medline devenue Pubmed, Inspec, Historical Abstracts, EconLit, Sociological Abstracts, etc.

La valeur scientifique de ces outils de repérage, d'analyse et d'homogénéisation de la documentation mondiale dans de nombreuses disciplines rend toujours un grand service aux chercheurs avancés. Les universités les mieux dotées assument encore des abonnements nombreux à ces outils<sup>176</sup> – et c'est un facteur d'attractivité des meilleurs étudiants et chercheurs. Les grandes universités de recherche proposent à leurs chercheurs l'accès à des bases comme Chemical Abstracts ou Historical Abstracts, qui apportent une exhaustivité dans leur discipline (y compris dans des langues ou des types de documents très divers – rapports techniques, plans, statistiques...) et une finesse dans les modes de recherche (on a mentionné la recherche par molécule dans CAS). Notons aussi que Chemical Abstracts assigne des CAS Number, identifiant de substances chimiques, et maintient un registre de ces substances, montrant la potentialité de ces services très spécialisés, au-delà de la recherche bibliographique.

Mais la connaissance de ces outils tend à disparaître parmi les nouvelles générations d'utilisateurs potentiels et la concurrence est forte avec les outils de découverte, les grands moteurs de recherche et les archives ouvertes.

Divers exemples montrent que le destin des bases bibliographiques spécialisées ou thématiques n'est pourtant pas scellé : certes, Pascal et Francis, produits par l'Inist de 1972 à 2015, ont été arrêtés, faute de rentabilité économique, mais peuvent être interrogés gratuitement sur le web dans leur version de 2015. Medline a conservé sa popularité, grâce à son passage au libre accès sur le web en 1996 et son enrichissement par l'accès aux publications ouvertes ; sa connexion à divers outils utiles aux biologistes moléculaires avec Pubmed lui ont permis de devenir le plus grand outil bibliographique mondial. Dans les domaines techniques, la plupart des outils ont évolué pour intégrer, outre les données bibliographiques exhaustives, des fonctionnalités de recherches avancées, l'analyse des citations, et surtout le lien au texte intégral (ACM, Inspec, IEEEExplore, SciFinder...).

Mais le paysage évolue encore, avec le développement mondial de l'Open Access et celui des bases ouvertes thématiques de preprints (BioRxiv, ChemRxiv, SocArXiv), alors que les données bibliographiques des principaux éditeurs sont de plus en plus accessibles librement via Crossref. La valeur ajoutée des bases payantes est remise en cause et les obligera à se renouveler encore et encore pour survivre.

---

<sup>176</sup> Voir par exemple les bases de données proposées par les Hautes écoles spécialisées suisses : <https://fher.swissuniversities.ch/fr/services/ressources-electroniques-hes/acces-par-type-de-ressource/bases-de-donnees/>. Consulté le 28 février 2021

### *Les bases de données de citations : Web of Science et Scopus*

Web of Science, base de données dont l'histoire remonte à 1964 et aujourd'hui détenue par Clarivate Analytics, et Scopus, base de données d'Elsevier créée en 2004, occupent une place centrale dans l'écosystème des bases de données de la recherche. Propriétaires, fermées et onéreuses, elles offrent pourtant un éventail de services de haut niveau qui les rendent attractives malgré leur coût, que tous les établissements ne peuvent cependant pas payer.

Outre les multiples fonctionnalités de recherche bibliographique, ces deux bases offrent des services bibliométriques encore incontournables dans l'évaluation de la recherche mondiale (WoS notamment est le pilote de l'Impact Factor, Scopus est leader en matière de h-index). Ces acteurs sont propriétaires d'une masse de données communiquées par des organismes de recherche ou des éditeurs, qu'ils traitent et enrichissent et qu'ils sont ensuite en position de monnayer, même si eux-mêmes les ont obtenues gratuitement. Par exemple, les données de citations sont notoirement riches dans ces bases, alors que des bases ouvertes (comme Crossref) peuvent avoir plus de mal à se les procurer (même si l'initiative I4OC a vocation à pallier cela).

Certains de leurs référentiels sont aussi d'une qualité inégalée – souvent grâce à un travail de curation fourni par les institutions clientes de ces bases –, ce qui peut amener certains établissements à en faire usage. Bibliolabs de l'université de Saclay, que nous présenterons au chapitre 5, s'appuie ainsi sur Scopus car leur référentiel d'affiliations est plus précis que ce qu'on peut trouver dans des bases ouvertes (ou dans WoS sur ce point). Mais ces données étant fermées, les données de Bibliolabs ne peuvent elles-mêmes pas être directement exposées de façon ouverte. Une initiative comme l'Open Science Monitor de la Commission européenne a également recours aux données de Scopus, et est critiquée pour cela<sup>177</sup>. Il y a donc un dilemme autour de l'usage de ces bases fermées : les utiliser revient à soutenir des initiatives antithétiques avec l'ouverture et le partage des données, mais les ignorer implique de se priver de données très utiles (tout en faisant des économies substantielles).

Aujourd'hui on peut se demander quel sera le destin de bases « citationnelles » payantes à l'heure du développement de bases offrant des services similaires gratuitement, comme Lens.org ou Dimensions.ai. Et alors que de nombreux grands éditeurs ont fini par accepter sous la pression de mettre à disposition librement leurs données de citation et leurs résumés via Crossref.

### *Les bases de données en accès libre*

Plus récemment sont apparus deux bases de données gratuites susceptibles de concurrencer les géants payants Scopus et WoS, voire Google Scholar : The Lens et Dimensions. Ces deux bases reposent sur les données d'autres acteurs, notamment pour les données de citations (qui viennent de Crossref et d'OpenCitations), mais elles développent des traitements intéressants de ces données. Elles offrent ainsi des fonctionnalités de recherche riches dans toutes les disciplines. The Lens est la propriété de l'organisation à but non lucratif Cambia, tandis que Dimensions est développé par Digital Science et repose sur un modèle freemium, notamment pour les fonctionnalités bibliométriques. En revanche, The Lens propose des services comparables à Scopus et

---

<sup>177</sup> « Open Metadata of Scholarly Publications », Ludo Waltman, 2019, [https://ec.europa.eu/info/sites/info/files/research\\_and\\_innovation/open\\_metadata\\_of\\_scholarly\\_publications\\_0.pdf](https://ec.europa.eu/info/sites/info/files/research_and_innovation/open_metadata_of_scholarly_publications_0.pdf). Consulté le 26 janvier 2021

WoS, mais entièrement gratuitement<sup>178</sup>. Il recueille des données de Crossref, PubMed, Core, Unpaywall, Orcid, mais surtout Microsoft Academic, qui lui permet de contenir plus de 228 millions de notices de travaux académiques, et 128 millions de brevets, sa spécialité<sup>179</sup>. Pour autant, comme l'estime Éric Jeangirard du Mesri, la couverture pour la France et les SHS de The Lens reste modeste, et un outil comme ScanR (dont nous reparlerons au chapitre 5) est plus complet sur le périmètre français.

### *Les moteurs de recherche académiques : Google Scholar et Microsoft Academic*

Google Scholar est probablement l'outil de recherche bibliographique le plus populaire dans les milieux de la recherche. Fondé sur les méthodes d'indexation automatique de Google (aux algorithmes non ouverts), donc sur le crawling de pages web et non sur la récupération de données par des API, il offre une masse de données (de l'ordre de près de 400 millions d'articles en 2019 selon certaines estimations<sup>180</sup> contre 70 millions dans Scopus en 2020 par exemple) et une rapidité de mise à jour inégalées. Sa puissance réside notamment dans sa capacité à indexer le plein-texte des articles, y compris derrière des *paywalls*, grâce à des accords avec la plupart des grands éditeurs<sup>181</sup>. Il propose un service populaire de profils de chercheurs et a servi de socle au développement d'un logiciel comme Publish or Perish, qui permet de recueillir et analyser des données de citations comme le h-index. Il est aussi moins sélectif que les bases précédemment citées, ce qui lui permet de recueillir des données dans des archives ouvertes par exemple, mais ce qui peut aussi l'amener à recueillir les données de revues de qualité faible (voire prédatrices). Ses fonctionnalités de recherche sont limitées et la précision de ses réponses est moyenne, rendant difficile une utilisation pour une revue systématique ou une méta-analyse par exemple. Par ailleurs, Google Scholar ne propose pas d'API : il est adapté à un usager qui souhaite trouver des articles pour sa propre recherche, et ce gratuitement, mais pas pour récupérer des données en masse ou mener des analyses bibliométriques poussées. Enfin, la pérennité du service, qui ne rapporte a priori pas grand-chose à Google, est dépendante à long terme du bon vouloir de cette multinationale.

Microsoft Academic, développé depuis 2016 par Microsoft Research, s'appuie sur le contenu indexé par Bing (le moteur de recherche de Microsoft) et lui applique des méthodes de traitement automatique des langues (*natural language processing*, NLP) pour en extraire de l'information, et des techniques d'inférence sémantique pour délivrer les résultats les plus pertinents pour l'utilisateur<sup>182</sup>. Contrairement à Google Scholar, Microsoft Academic n'explore pas le plein-texte, mais cela lui permet (en l'absence de données sous droits) de rendre ses données disponibles via le Microsoft Academic Graph en open data (sous licence ODC-BY)<sup>183</sup>, et accessible via une API. Ainsi, quoique le

<sup>178</sup> « 7 reasons why you should try Lens.org », Aaron Tay, Medium, 2018, <https://aaron.tay.medium.com/6-reasons-why-you-should-try-lens-org-c40abb09ec6f>. Consulté le 18 février 2021

<sup>179</sup> Page d'accueil, The Lens, <https://www.lens.org/>. Consulté le 3 mars 2021

<sup>180</sup> Selon une estimation de Michael Gusenbauer dans « Google Scholar to Overshadow Them All? Comparing the Sizes of 12 Academic Search Engines and Bibliographic Databases », *Scientometrics* n° 118/1, 2019, <https://doi.org/10.1007/s11192-018-2958-5>. Consulté le 21 février 2021

<sup>181</sup> « The next Generation Discovery Citation Indexes – A Review of the Landscape a 2020 (I) », Aaron Tay, Medium, 2020, <https://medium.com/a-academic-librarians-thoughts-on-open-access/the-next-generation-discovery-citation-indexes-a-review-of-the-landscape-a-2020-i-afc7b23ceb32>. Consulté le 18 février 2021

<sup>182</sup> « How is MA different from other academic search engines? », Microsoft Academic, <https://academic.microsoft.com/faq>. Consulté le 22 février 2021

<sup>183</sup> « The next Generation Discovery Citation Indexes — A Review of the Landscape a 2020 (I) », Aaron Tay, Medium, 2020, <https://medium.com/a-academic-librarians-thoughts-on-open-access/the-next-generation-discovery-citation-indexes-a-review-of-the-landscape-a-2020-i-afc7b23ceb32>. Consulté le 18 février 2021

volume de données exposé soit moindre que celui de Google Scholar, Microsoft Academic est un acteur potentiellement majeur des métadonnées ouvertes<sup>184</sup>. Cependant, comme Google Scholar, son avenir dépend de la multinationale dont il dépend.

### *Les index d'outils de découverte*

Les outils de découverte permettent d'agréger des données d'éditeurs, de bases de données externes et des données locales (catalogues de bibliothèques, archives institutionnelles...). Il s'agit d'outils le plus souvent propriétaires, même si des solutions open source existent, mais sont peu utilisées en France<sup>185</sup>. Les principaux concurrents sont Primo Central d'ExLibris, Ebsco Discovery Service et WorldCat Discovery d'OCLC. Ils sont commercialisés notamment auprès des bibliothèques.

Les index de ces outils de découverte sont énormes mais la qualité des métadonnées n'y est souvent pas optimale, en raison de la masse de données de sources multiples, et souvent hétérogènes, qu'ils ont à traiter et exposer. Le souci d'efficacité peut aussi amener à appauvrir des métadonnées riches au cours d'opérations d'alignements ou de débouclage<sup>186</sup>. Cela peut occasionner par exemple des problèmes de résolution de liens vers les ressources électroniques.

En conséquence, et malgré leur coût important pour les établissements, les outils de découverte ne sont parfois pas très populaires auprès des usagers, qui peuvent leur préférer des outils comme Google Scholar.

### *Les archives ouvertes et leurs moissonneurs*

Les archives ouvertes constituent des bases de données particulièrement importantes dans le cadre de la science ouverte. ArXiv est la première archive ouverte de l'histoire, lancée en 1991 par Paul Ginsparg, qui propose des prépublications et des postpublications d'articles dans les domaines de la physique, des mathématiques, de l'informatique, de la biologie quantitative, de la finance quantitative, de la statistique, de l'ingénierie électrique et des systèmes, et de l'économie. Aujourd'hui, il en existe plus de 5 600 recensées par le répertoire OpenDoar du Jisc<sup>187</sup> : à l'échelle d'un établissement (archives institutionnelles) ou d'un pays (HAL en France est cependant un exemple rare), mondiales et transdisciplinaires (Zenodo) ou disciplinaires (bioRxiv, PubMed Central, Repec...). Elles peuvent être alimentées par auto-archivage ou versements par les éditeurs.

Le protocole OAI-PMH permet le moissonnage des données vers d'autres bases de données, comme BASE (Bielefeld Academic Search Engine), qui recense plus de 240 millions de documents de plus de 8 000 sources, mais ne contient pas de plein-texte. Un autre exemple est la base britannique CORE (COnnecting REpositories), qui contient près de 210 millions de documents et leurs métadonnées, moissonnées d'archives ouvertes et

---

<sup>184</sup> « Open Metadata of Scholarly Publications », Ludo Waltman, 2019, [https://ec.europa.eu/info/sites/info/files/research\\_and\\_innovation/open\\_metadata\\_of\\_scholarly\\_publications\\_0.pdf](https://ec.europa.eu/info/sites/info/files/research_and_innovation/open_metadata_of_scholarly_publications_0.pdf). Consulté le 26 janvier 2021

<sup>185</sup> « Les outils de découverte en bibliothèque universitaire », Soledad Lida, Enssib, 2016, <https://www.enssib.fr/bibliotheque-numerique/documents/67305-les-outils-de-decouverte-en-bibliotheque-universitaire.pdf>. Consulté le 19 février 2021

<sup>186</sup> Ibid.

<sup>187</sup> « OpenDoar Statistics », OpenDoar, [https://v2.sherpa.ac.uk/view/repository\\_visualisations/1.html](https://v2.sherpa.ac.uk/view/repository_visualisations/1.html). Consulté le 19 février 2021

d'éditeurs en accès ouvert ou hybrides<sup>188</sup>.

### *Vers un graphe mondial des publications scientifiques ?*

Le cadre du web sémantique permet des alignements multiples entre les données, constituant *in fine* un gigantesque graphe reliant potentiellement l'intégralité des entités et des informations portant sur elles, au moyen de relations sémantiques (de type « est l'auteur de », « est cité par », etc.). Dans l'univers des bibliothèques, comme le souligne Françoise Leresche<sup>189</sup>, de la BnF, l'enjeu de la Transition bibliographique est ainsi de faire des catalogues de bibliothèques et des bases bibliographiques en général des graphes alignés et liés à d'autres graphes, pour être visible sur le web, grâce aux données d'autorité.

Openaire et son Openaire Research Graph<sup>190</sup> constituent une initiative ouverte (les données sont en CC-BY ou CC0) pour récolter les métadonnées de quelque 10 000 sources « *de confiance* » (archives ouvertes, éditeurs, registres, agrégateurs, etc.)<sup>191</sup>, qu'il complète par la fouille de 12 millions d'article Open Access en plein texte. En février 2021, le Graphe revendiquait signaler 125 millions de publications, 14 millions de jeux de données et 220 000 logiciels de recherche.

Semantic Scholar<sup>192</sup>, développé par l'Allen Institute, utilise les technologies de l'intelligence artificielle pour analyser le contenu sémantique des publications scientifiques et connecter de façon fine les multiples productions scientifiques entre elles<sup>193</sup>. En 2019, son corpus incluait plus de 180 millions de publications<sup>194</sup>.

À l'échelle mondiale, le projet le plus abouti et le meilleur candidat à ce stade pour jouer le rôle de réservoir libre de métadonnées, y compris des publications scientifiques, semble être Wikidata.

Selon le Comité pour la science ouverte, « depuis 2012, la base Wikidata est devenue progressivement le point de convergence mondial des identifiants ouverts »<sup>195</sup>. Wikidata est ainsi un nœud permettant de faire des alignements entre identifiants et de connecter les entités entre elles, au-delà de l'univers des bibliothèques et des publications scientifiques. Il propose une interface simple de gestion des données ouvertes liées aux établissements universitaires qui n'auraient pas les moyens techniques de développer une

<sup>188</sup> « About Core », Core, <https://core.ac.uk/about/>. Consulté le 22 février 2021

<sup>189</sup> Dans la vidéo « N'oubliez pas les données d'autorité », 4e journée du Groupe Systèmes & Données du programme Transition bibliographique, 2019, [https://www.youtube.com/watch?time\\_continue=18000&v=y1qc3veB-wM&feature=emb\\_logo](https://www.youtube.com/watch?time_continue=18000&v=y1qc3veB-wM&feature=emb_logo), vers 22". Consulté le 29 janvier 2021

<sup>190</sup> « The OpenAIRE Research Graph », Openaire, <https://www.openaire.eu/blogs/the-openaire-research-graph>. Consulté le 29 janvier 2021

<sup>191</sup> « OpenAIRE data sources are considered "trusted" when researchers rely on them to share, discover, monitor, and assess their scientific products. Known sources collected by OpenAIRE are institutional repositories (e.g. university archives, libraries), catch-all repositories (e.g. Zenodo, Figshare, Dryad, B2Share, etc.), data repositories (e.g. Pangaea, GESIS, bio sources), thematic repositories (e.g. ArXiv, EuropePMC, RePec, etc.), Open Access publishers (e.g. F1000, OpenEdition, etc.), knowledge graphs (e.g. Microsoft Academic Graph, OpenCitations), registries (e.g. CrossRef, DataCite, ORCID, GRID.ac, OpenDOAR, re3data.org, etc), aggregators (e.g. Unpaywall, BASE, Scielo, DOAJ, CORE-UK, etc.), funders (e.g. European Commission, NSF, Wellcome Trust, etc.) », voir sur <https://www.openaire.eu/blogs/the-openaire-research-graph>. Consulté le 29 janvier 2021

<sup>192</sup> Page d'accueil, Semantic Scholar Research, <https://research.semanticscholar.org/>

<sup>193</sup> Voir notamment Kyle Lo et al., « S2ORC: The Semantic Scholar Open Research Corpus », in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020, Online: Association for Computational Linguistics, 2020), p. 4969-83, <https://doi.org/10.18653/v1/2020.acl-main.447>. Consulté le 28 février 2021

<sup>194</sup> « About us », Semantic Scholar, <https://pages.semanticscholar.org/about-us>. Consulté le 18 février 2021

<sup>195</sup> « Des identifiants ouverts pour la science ouverte : note d'orientation », Comité pour la science ouverte, 2019, <https://www.ouvrirlascience.fr/des-identifiants-ouverts-pour-la-science-ouverte-note-dorientation/>. Consulté le 5 mars 2021



infrastructure propre<sup>196</sup>.

Initiative totalement ouverte, appuyée sur la Wikimedia Foundation, elle est une bonne candidate pour le statut de hub mondial centralisant et alignant les identifiants, les données d'autorités et les données bibliographiques. La BnF verse ainsi ses données d'autorité dans Wikidata depuis plusieurs années, mais des établissements plus modestes peuvent aussi y verser simplement des données, à des échelles plus modestes, par exemple avec l'outil OpenRefine<sup>197</sup>. Le CoSO fait figurer parmi ses objectifs l'alignement des identifiants chercheurs français avec Wikidata. Notons qu'un groupe de travail de l'Ifla a été constitué fin 2019 pour œuvrer à l'implication des bibliothèques dans Wikidata et au développement d'alignements avec les formats Marc, RDA et Bibframe<sup>198</sup>.

Wikidata contient notamment des données (d'autorité, bibliographiques, et de citations, en lien avec le projet Wikicite<sup>199</sup>) sur les productions scientifiques. Ces données peuvent être exploitées avec des outils comme Scholia<sup>200</sup>, à travers l'interrogation en Sparql du Wikidata Query Service. Son code est disponible sur Github<sup>201</sup>. Scholia<sup>202</sup> permet de créer des fiches sur des chercheurs, des thèmes de recherche, des institutions, des revues, des projets ou même des molécules chimiques et des espèces, etc., et d'afficher des bibliographies, des visualisations de données, de statistiques, etc. L'outil permet de prendre la mesure des données présentes dans Wikidata. Pour autant le versement de données par davantage d'acteurs dans le monde académique permettrait d'aller plus loin encore, et de limiter les biais et l'incomplétude dont souffre encore Wikidata.

## Bilan d'étape

Dans ce deuxième chapitre, nous avons décrit l'écosystème des acteurs et producteurs des métadonnées des publications, et leur importance pour la science ouverte. L'architecture de cet écosystème étant complexe et en constante évolution, sa compréhension nécessite un travail de formation et de veille. Mais cette connaissance fait partie du bagage nécessaire des professionnels de l'IST investis sur la question des données et métadonnées académiques, afin d'être en mesure d'adopter une approche stratégique et non simplement opérationnelle. Penser globalement permet, nous semble-t-il, de mieux agir localement, et d'agir à la fois pour les intérêts propres de son

<sup>196</sup> « Wikidata: a platform for your library's linked open data », Stacy Allison-Cassin et Dan Scott, *The Code4Lib Journal*, 2018, <https://journal.code4lib.org/articles/13424>. Consulté le 16 février 2021

<sup>197</sup> Un exemple dans un établissement universitaire français est celui de la bibliothèque interuniversitaire de santé (Paris Sorbonne) qui, en 2018, dans le cadre d'un partenariat avec Wikimedia pour les versements de portraits de leur bibliothèque numérique, ont décidé de lier les métadonnées d'autorité des images avec Wikidata. À l'époque, OpenRefine ne proposant pas encore de connexion directe avec Wikidata, il avait été décidé de passer par Viaf : l'opération a consisté à d'abord aligner les données de leur base avec Viaf, pour ensuite, à partir d'un dump des données Viaf, elles-mêmes alignées sur Wikidata, aligner leurs données sur celles de Wikidata. Plus récemment, un alignement avec Wikidata a été effectué pour des images d'établissements de santé, pour ces établissements et les villes où elles se trouvent : cette fois-ci, OpenRefine proposait un alignement direct avec Wikidata, rendant l'opération plus simple. Toutefois, le maniement d'un outil comme OpenRefine n'est pas intuitif et demande une formation préalable, qui peut parfois manquer dans certains établissements. C'est cependant un investissement intéressant dans la mesure où l'outil a un grand nombre d'usages dans le domaine de la curation de données en masse. Voir « Les portraits de la BIU Santé dans Wikimedia Commons », Le blog actualités de la BIU Santé, 3 juillet 2018, <https://www.biusante.parisdescartes.fr/blog/index.php/portraits-de-biu-sante-wikimedia-commons/>. Consulté le 11 février 2021

<sup>198</sup> « Ifla Wikidata Working Group », Ifla, <https://www.ifla.org/node/92837>. Consulté le 16 février 2021

<sup>199</sup> « WikiCite », Wikimedia, <https://meta.wikimedia.org/wiki/WikiCite>. Consulté le 22 février 2021

<sup>200</sup> « Scholia, Scientometrics and Wikidata », Finn Årup Nielsen, Daniel Mietchen, et Egon Willighagen, in *The Semantic Web: ESWC 2017 Satellite Events*, éd. par Eva Blomqvist et al., *Lecture Notes in Computer Science* (Cham: Springer International Publishing, 2017), p. 237-59, [https://doi.org/10.1007/978-3-319-70407-4\\_36](https://doi.org/10.1007/978-3-319-70407-4_36). Consulté le 14 février 2021

<sup>201</sup> Code de Scholia, Github, <https://github.com/fnielsen/scholia>; Consulté le 14 février 2021

<sup>202</sup> Page d'accueil, Scholia, <https://scholia.toolforge.org/>. Par exemple, ici, la fiche du linguiste américain Noam Chomsky : <https://scholia.toolforge.org/author/Q9049>. Consultés le 14 février 2021

## **Chapitre 2 : gouvernance de la science ouverte et des métadonnées de la recherche**

établissement ou de sa communauté de recherche, et pour l'intérêt commun, en œuvrant à la construction d'outils partagés et ouverts. Notamment, une telle connaissance peut permettre de mieux se positionner vis-à-vis de l'administration des identifiants, de la structuration des référentiels de la recherche et de la gestion des métadonnées des éditeurs, que nous allons examiner dans la deuxième partie de ce mémoire.

## **PARTIE B : LA GESTION DES MÉTADONNÉES DANS LES ÉTABLISSEMENTS DE L'ESR**

---

Dans cette seconde partie, nous nous pencherons sur les pratiques concrètes des bibliothécaires dans les établissements de l'ESR (bibliothèques universitaires et agences bibliographiques notamment) permettant d'œuvrer à des données et métadonnées plus ouvertes et de meilleure qualité. Nous le ferons selon trois axes : dans un premier temps, nous nous pencherons sur l'action que les professionnels de l'IST peuvent avoir en faveur des identifiants et référentiels, puis nous verrons comment il est possible d'intervenir pour améliorer les métadonnées des productions des institutions de recherche (dans le cadre d'archives ouvertes, de revues en accès ouvert), enfin nous étudierons les outils de traitements des données, notamment de façon automatisée, que ce soit à des fins bibliométriques ou pour construire des bases de données ouvertes de meilleure qualité.

### **CHAPITRE 3 : L'ENJEU DES IDENTIFIANTS ET RÉFÉRENTIELS**

Le cadre du web de données liées implique de sortir d'une logique de documents pour aller vers une logique d'entités. À ce titre, les identifiants occupent une place centrale, car ils permettent d'identifier de manière unique et pérenne des entités (publications, chercheurs, institutions de recherche, concepts...), et peuvent donc servir de pivot, de passerelle entre les données.

#### **Identifiants pérennes et interopérabilité sur le web sémantique**

Les identifiants pérennes<sup>203</sup> rendent possible l'interopérabilité. Comme le souligne Françoise Leresche<sup>204</sup>, de la BnF, les identifiants internationaux normalisés notamment servent de pivot à l'alignement entre les systèmes locaux. En leur absence, il faut produire un alignement spécifique avec chaque système séparément : par exemple, les noms de lieux n'ayant pas d'identifiant international normalisé de référence, il faut associer une notice de lieu de la BnF avec son correspondant dans DBpedia, Wikidata, Geonames, Insee... À l'inverse, une notice de personnes peut être directement associée à un Isni, et sera ainsi alignée automatiquement avec un large ensemble de systèmes dans le monde. On peut ainsi aligner entre eux l'Isni, l'Orcid, l'IdRef, l'IdHAL, le Vial, le Q Wikidata, des identifiants de bibliothèques nationales, etc. Un identifiant auteur permet d'associer au chercheur ses publications, si un identifiant a été indiqué dans les métadonnées de ces

<sup>203</sup> Voir le chapitre 2 pour une description de quelques institutions gérant des identifiants (Crossref, DataCite, ORCID...).

<sup>204</sup> Dans la vidéo « N'oubliez pas les données d'autorité », 2019, [https://www.youtube.com/watch?time\\_continue=18000&v=y1qc3veB-wM&feature=emb\\_logo](https://www.youtube.com/watch?time_continue=18000&v=y1qc3veB-wM&feature=emb_logo), vers 19'. Consulté le 19 janvier 2021

ressources. Les identifiants permettent également de résoudre des problèmes d'homonymie ou au contraire de variations dans les formes du nom d'une personne, par exemple dans différentes langues ou différents alphabets. Les identifiants permettent enfin de tracer des ponts entre des ressources de domaines différents, en dehors des bibliothèques, entre des environnements techniques différents (Marc et non-Marc par exemple)<sup>205</sup>.

Le Comité pour la science ouverte (CoSO) définit l'identifiant comme « un numéro ou une étiquette alphanumérique, opaque ou explicite, lisible par des machines et par des humains, permettant de désigner et de retrouver de manière univoque et pérenne un objet, un document, une personne, un lieu, un organisme, ou toute entité, dans le monde physique ou numérique » et « souvent associé à une URI<sup>206</sup> »<sup>207</sup>.

Il existe des identifiants pour de très nombreuses entités, tous n'étant pas pertinents pour le monde de la recherche et n'ayant pas le même degré de maturité. Comme le souligne Freya<sup>208</sup>, programme de l'Union européenne pour le développement des identifiants pérennes, seuls les identifiants pour les chercheurs, les publications et les données peuvent être considérés comme matures à ce jour, mais de nombreux acteurs s'efforcent de les développer pour de nombreuses autres entités (organisations, financements, logiciels, conférences, etc.)<sup>209</sup>.

Les publications scientifiques sont représentés notamment par les DOI, mais peuvent aussi l'être par Handle ou, dans des bases spécifiques, PubMed Central (PMC) ID, arXiv ID, etc. Notons qu'un DOI peut aussi identifier des conférences, des subventions, des jeux de données. Les identifiants ARK (Archival Resource Key), utilisés surtout par des institutions patrimoniales comme les bibliothèques ou les musées, ont été créés en 2001 par la California Digital Library, et permettent d'identifier tout type de ressources, physiques, numériques, immatériels... Cela permet par exemple à la BnF d'identifier tant leurs notices de catalogue que leurs ressources numériques. À un niveau de granularité moins fin, les identifiants ISBN et ISSN permettent d'identifier des ouvrages et des revues<sup>210</sup>.

Les identifiants de personnes permettent d'identifier les chercheurs, notamment par l'Isni (International Standard Name Identifiers) et l'Orcid (Open Researcher and Contributor Ids), mais aussi le Vial (Virtual International Authority Files...). Là encore, il existe également des identifiants propriétaires de chercheurs, comme ResearchID (Clarivate Analytics, qui gère notamment le Web of Science et le Facteur d'impact) et Scopus ID (de la base de données Elsevier Scopus), et des identifiants d'archives ouvertes comme arXiv ou IdHAL, mais ces identifiants ont un rôle par définition plus local, et ils doivent donc être alignés sur des identifiants de référence (comme IdRef pour IdHAL).

<sup>205</sup> « Transitioning to the Next Generation of Metadata », Karen Smith-Yoshimura, 2020, <https://www.oclc.org/content/dam/research/publications/2020/oclcresearch-transitioning-next-generation-metadata-a4.pdf>. Consulté le 29 janvier 2021

<sup>206</sup> L'URI (Uniform Resource Identifier), standard du web sémantique, permet d'identifier une ressource sur Internet, même si elle est supprimée ou déplacée.

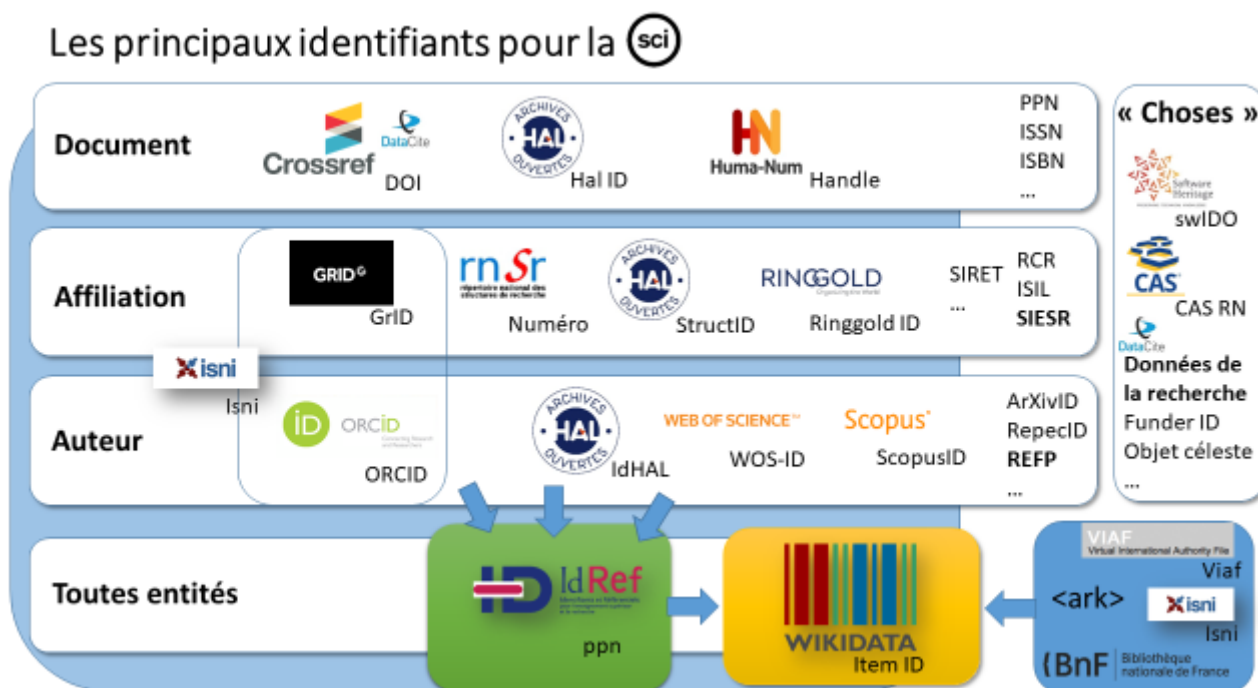
<sup>207</sup> « Des identifiants ouverts pour la science ouverte : note d'orientation », Comité pour la science ouverte, 2019, <https://www.ouvrirlascience.fr/des-identifiants-ouverts-pour-la-science-ouverte-note-dorientation/>. Consulté le 29 janvier 2021

<sup>208</sup> « The Freya project », Freya, <https://www.project-freya.eu/en/about/mission>. Consulté le 29 janvier 2021

<sup>209</sup> « Survey of Current PID Services Landscape », Christine Ferguson, Jo McEntyre (EMBL-EBI) Vasily Bunakov, Simon Lambert (STFC) Stephanie van der Sandt (CERN) Rachael Kotarski, Sarah Stewart, Andrew MacEwan (BL) Martin Fenner, Patricia Cruse (DataCite) René van Horik (DANS) Tina Dohna, Ketil Koop-Jacobsen, Uwe Schindler (PANGAEA) Siobhan McCafferty (ANDS) / Freya, 2018, [https://www.project-freya.eu/en/deliverables/freya\\_d3-1.pdf](https://www.project-freya.eu/en/deliverables/freya_d3-1.pdf). Consulté le 29 janvier 2021. On trouvera dans ce document un tableau récapitulatif de l'ensemble de types d'identifiants à un stade mature, émergent ou immature.

<sup>210</sup> Cependant, les ISBN sont parfois réattribués, et ne peuvent donc pas être considérés comme durable.

Le schéma suivant, proposé par le Comité pour la science ouverte<sup>211</sup>, montre les différents identifiants utiles pour la science et leurs relations, suivant les types d'entités :



Les principaux identifiants pour les documents, les affiliations, les auteurs, les « choses » et toutes entités<sup>212</sup>

## Quels critères d'identifiants de qualité ?

Les identifiants doivent respecter un ensemble de règles pour jouer leur rôle de façon optimale. Le Comité pour la science ouverte (CoSO) définit les critères suivants d'identifiants de qualité pour la science ouverte :

- précis (1 identifiant correspond à 1 entité)
- durables (1 identifiant est attribué pour toujours à une entité et n'est pas réattribué à 1 autre entité si la première disparaît)
- ouverts et interopérables (trouvables, échangeables, et utilisables par d'autres registres, par des humains comme par des machines, à l'unité ou par lots, sans abonnement ni inscription préalable)
- documentés (règles d'attribution définies, publiées, évolutives si besoin, fruit du consensus d'une communauté)
- réconciliables (possibilité pour les producteurs des registres ou pour des tiers de relier les identifiants d'une même entité présente dans plusieurs registres)
- régis par un système transparent de gouvernance. »<sup>213</sup>

Ces critères répondent, on le voit, aux principes Fair présentés plus haut (chapitre

<sup>211</sup> « Des identifiants ouverts pour la science ouverte : note d'orientation », Comité pour la science ouverte, 2019, <https://www.ouvri.lascience.fr/des-identifiants-ouverts-pour-la-science-ouverte-note-dorientation/>. Consulté le 29 janvier 2021

<sup>212</sup> Crédit : Comité pour la science ouverte

<sup>213</sup> Ibid.

1). On retrouve les exigences de standardisation, de lisibilité par les machines, de partage, d'ouverture. Tous ces critères sont bien sûr interdépendants : un identifiant ne peut être pérenne que s'il s'appuie sur un système solide de gouvernance, soutenu par un modèle économique durable ; il ne peut être partagé que s'il est interopérable et ouvert.

## Identifiants et référentiels

Isolé, sans interconnexion, un identifiant perd une large partie de sa puissance. C'est pourquoi le CoSO distingue deux types de services sur les identifiants :

« • Les services permettant la création d'identifiants pérennes standards pour divers types d'entités, disponibles à chaque instant dans l'environnement de chaque utilisateur (chercheur, administrateur) ;

• Les services permettant la réconciliation des identifiants pérennes et l'interconnexion des registres à l'échelle internationale, ainsi que l'exposition libre des données. Ces services sont appelés **référentiels**. »<sup>214</sup>

Un identifiant doit être diffusé, aligné, et pour cela apparaître dans un référentiel partagé, qui peut associer des entités à leurs divers identifiants et créer du lien entre les identifiants et donc entre les entités elles-mêmes. Par exemple, un référentiel peut permettre de lier un chercheur à ses identifiants Orcid, IdRef, IdHAL, ScopusId, puis lier le chercheur à ses publications, elles-mêmes identifiées par des DOI, et à ses institutions de rattachement identifiées par IdRef, Grid, etc. Selon François Mistral, de l'Abes, « un référentiel est un jeu de données, suffisamment vraies, justes, certaines pour être utilisées en confiance afin d'en produire ou d'en agréger d'autres. De fait, ces données de référence sont des points de repère à partir desquelles en situer d'autres avec économie »<sup>215</sup>. Un référentiel doit permettre de gagner en efficacité, et manque son but si ses utilisateurs doivent effectuer un travail disproportionné de contrôle et de traitement pour pouvoir l'utiliser sans risque d'erreurs. Un référentiel doit donc faire référence.

On peut distinguer à ce titre les référentiels des « registres », ces derniers ne bénéficiant pas d'une curation centralisée ou coordonnée et permettant l'autodéclaration d'entités (comme dans Orcid ou IdHAL, où c'est le chercheur qui crée lui-même son identifiant). Cette pratique peut engendrer des problèmes de qualité, avec le risque supplémentaire, dans le cadre du web de données, que les erreurs se diffusent à large échelle, en raison même de ce qui fait la force du web sémantique, la liaison entre les données. On ne saurait donc trop insister sur la nécessité pour les professionnels de l'IST de maintenir de très hauts standards de qualité sur les référentiels dont ils ont la responsabilité, à quelque niveau que ce soit.

Cependant, les bibliothécaires sont bien placés pour effectuer ce travail. Ils disposent d'emblée de compétences permettant de gérer les identifiants, et donc les référentiels, car un référentiel est proche d'un fichier d'autorité, même si c'est dans un cadre élargi et différent de celui d'un catalogue.

Deux référentiels sont particulièrement importants pour l'ESR français :

- **IdRef** (Identifiants et référentiels pour l'enseignement supérieur et la recherche) est une « application web développée et maintenue par l'Abes.

<sup>214</sup> Ibid.

<sup>215</sup> « Autorités vs référentiels : 3 questions aux experts de l'Abes », Punktokomo, 2017, <https://punktokomo.abes.fr/2017/04/20/autorites-vs-referentiels-3-questions-aux-experts-de-labes/>

IdRef permet à des utilisateurs et à des applications tierces d'interroger, de consulter, de créer et d'enrichir des autorités »<sup>216</sup>. IdRef est fondé sur les autorités Sudoc, et donc sur un réseau de catalogueurs experts, qui assurent une curation coordonnée (même si elle est décentralisée à travers tout l'ESR français). IdRef les rend requêtables par le moteur de recherche Solr et moissonnables en OAI-PMH, ainsi qu'en RDF pour une diffusion sur le web de données. Cette infrastructure porte un effort particulier sur la question de l'interopérabilité et les alignements : « Grâce à un programme dédié, il est désormais possible d'aligner sur IdRef – ainsi que sur d'autres référentiels tels Isni, Vial, Orcid – des entités “personnes” issues de différentes sources (catalogues documentaires, archives institutionnelles, entrepôts OAI-PMH, bases d'articles, annuaires, référentiels eux-mêmes...). À ce jour, l'alignement à IdRef de 50 000 autorités Persée constitue le plus bel exemple. »<sup>217</sup> De nombreux projets disciplinaires, menés dans les établissements, notamment via les appels à projet Collex-Persée, comme Datapoc (MNHN), Mistara (Bulac), Refdivinités (BIS)..., sont destinés à l'alignement de données d'autorité notamment avec IdRef<sup>218</sup> (voir plus bas). Ainsi, IdRef est un bon candidat pour jouer le rôle de référentiel national de l'ESR.

- Un autre référentiel (ou plutôt « registre », en l'absence de curation coordonnée) important pour l'ESR français est, au sein de HAL, Auréhal, « l'application de gestion des six référentiels utilisés dans HAL : domaines, structures, revues, projets ANR, projets européens et auteurs »<sup>219</sup>. HAL étant lui-même un bon candidat pour centraliser les productions scientifiques et leurs métadonnées à l'échelle nationale, le référentiel sur lequel il s'appuie est un enjeu crucial. Or Auréhal nécessite un gros travail de curation de la part des professionnels dans les établissements car les déposants ont la possibilité de créer des entités, ce qui occasionne des problèmes de qualité et de doublonnage. Notons qu'une archive institutionnelle comme Oatao a fait le choix de s'intégrer davantage dans IdRef, un choix que HAL se refuse pour le moment à faire, malgré les problèmes rencontrés avec Auréhal. En revanche, Auréhal a vocation à être aligné sur IdRef et Orcid.

## Quelles actions sont envisagées en France pour développer l'usage des identifiants et référentiels ?

Les objectifs stratégiques à l'échelle française proposés par le Comité pour la science ouverte sur les identifiants ouverts sont les suivants :

- « 1. accélérer l'adoption des identifiants par les chercheurs, les laboratoires et les institutions, afin de rendre plus visible la production scientifique française courante et cumulée ;
2. identifier des modèles économiques viables pour assurer un déploiement

<sup>216</sup> « A propos d'IdRef », IdRef, [https://www.idref.fr/a\\_propos.jsp](https://www.idref.fr/a_propos.jsp). Consulté le 29 janvier 2021

<sup>217</sup> « Qualinca et IdRef : l'intégration est en cours ! », Aline Le Provost, *Arabesques* n° 85, 2017, <https://publications-prairial.fr/arabesques/index.php?id=215>. Consulté le 29 janvier 2021

<sup>218</sup> « Alignements IdRef en soutien aux projets Collex-Persée des établissements (Les Actus de l'Abes - partie 1) », Abes, 2020, <https://vimeo.com/415091624>. Consulté le 1<sup>er</sup> février 2021

<sup>219</sup> « Auréhal et son IdHAL rassembleur », Bruno Marmol et Bénédicte Kuntziger, *Arabesques* n° 85, 2017, <https://publications-prairial.fr/arabesques/index.php?id=229>. Consulté le 1<sup>er</sup> février 2021

durable des identifiants à l'échelle nationale ;

3. améliorer l'interopérabilité et la normalisation des identifiants tout en s'assurant d'un contrôle par la communauté scientifique ;

4. contribuer au pilotage et à l'évolution des systèmes d'identifiants afin d'en garantir l'ouverture et l'indépendance sur le long terme »<sup>220</sup>.

On voit à travers ces préconisations que la problématique des identifiants ne se réduit ni à une simple question technique, ni à une question politique, ni à une question économique mais mêle ces différentes dimensions, qui bien sûr interagissent entre elles. La politique des identifiants aux échelles nationale et locale en France doit donc aller dans plusieurs directions : vers une meilleure qualité des identifiants et référentiels, vers une plus large adoption par les utilisateurs (chercheurs, structures de recherche...), et vers une meilleure mutualisation et de meilleurs alignements entre référentiels, y compris à l'échelle internationale.

Le travail des professionnels de l'IST est indispensable sur tous ces plans, y compris à l'échelle locale. En l'occurrence, les bibliothèques universitaires disposent de professionnels compétents sur ces sujets, notamment les correspondants autorités<sup>221</sup>, les correspondants Orcid<sup>222</sup>, et les responsables Science ouverte, Archives ouvertes ou Soutien à la recherche, qui peuvent œuvrer d'une part à la plus large adoption des identifiants par les chercheurs et leurs structures, d'autre part à la qualité des identifiants dans les notices de la communauté dont il s'occupe, permettant ensuite un travail d'alignement plus efficace par des institutions nationales comme l'Abes.

### *Inciter à l'adoption d'identifiants auteurs*

Une des tâches des professionnels de l'IST vis-à-vis des identifiants, notamment les identifiants auteurs, est d'œuvrer à ce que les utilisateurs potentiels les adoptent effectivement. Il s'agit d'une des multiples tâches qu'un service de soutien aux chercheurs au sein d'un SCD peut accomplir auprès des chercheurs de son établissement. Un identifiant comme Orcid étant créé par le chercheur lui-même, il est difficile de se substituer à lui dans cette démarche. Il est cependant possible d'apporter au chercheur un service de conseil, d'aide, de formation (qui peut prendre la forme d'un simple tutoriel ou de sessions en présentiel) pour la création d'un identifiant Orcid ou IdHAL. Il peut s'agir, en premier lieu, d'informer les chercheurs sur l'existence des identifiants et de clarifier leur rôle, de souligner l'importance cruciale des identifiants dans l'architecture des publications scientifiques actuelles, et d'autre part de montrer la procédure à suivre pour la création elle-même et pour d'autres fonctionnalités, alignements ou versements de références de publications d'une base de données à une autre (de HAL à Orcid par exemple)<sup>223</sup>.

Un travail auprès des chercheurs peut par exemple consister à montrer l'importance cruciale des identifiants pérennes dans le cadre de l'identité numérique, face à l'utilisation plus massive des réseaux sociaux académiques tels que Researchgate et Academia. À cet

<sup>220</sup> « Des identifiants ouverts pour la science ouverte : note d'orientation », Comité pour la science ouverte, 2019. <https://www.ouvri.la-science.fr/des-identifiants-ouverts-pour-la-science-ouverte-note-dorientation/>. Consulté le 1<sup>er</sup> février 2021

<sup>221</sup> « Les Correspondants Autorités », Abes, <https://abes.fr/reseaux-idref-orcid/le-reseau/correspondants-autorites/>. Consulté le 1<sup>er</sup> février 2021

<sup>222</sup> « Les Correspondants Orcid », Abes, <https://abes.fr/reseaux-idref-orcid/le-reseau/correspondants-orcid/>. Consulté le 1<sup>er</sup> février 2021

<sup>223</sup> Des formations aux formateurs sur l'identité numérique sont proposées par les Urlist par exemple : « L'identité numérique du chercheur : quel accompagnement ? », Aline Bouchard, 2018, <https://urfistinfo.hypotheses.org/3219>. Consulté le 1<sup>er</sup> février 2021



égard, une étude sur les chercheurs de l'université de Caen est éloquent : sur 619 chercheurs en STM, 416 étaient sur Researchgate et 157 avaient un Orcid, et la répartition était même de 153 contre 22 en SHS, sur un total de 428 chercheurs<sup>224</sup>. La feuille de route 2016-2020 du CCSD dresse une critique de ces réseaux : « Des initiatives généralement privées sont venues occuper le créneau des services offerts aux chercheurs autour des publications, les “réseaux sociaux scientifiques”. Ces services sont utiles aux chercheurs, à la mesure exacte de leur usage. Mais ils interfèrent avec le développement de “l'open science” en détournant la volonté réelle de partager les résultats de la recherche, partage qui se trouve limité aux collègues proches et prêts eux aussi à offrir leurs données personnelles, et s'exposant au risque permanent d'être (re-)privatisé. » Ainsi, de nombreux chercheurs donnent à des opérateurs privés des données qu'ils n'exposent pas sur des bases de données ouvertes. Le chercheur peut parfois aller au plus simple et au plus attractif, ce qui de son point de vue et à court terme pourra consister à créer une page et à diffuser ses articles (parfois illégalement) sur les réseaux sociaux académiques. Un argumentaire peut donc être déployé pour qu'il adopte une démarche de développement de ses identifiants pérennes et de dépôt de ses articles dans des archives ouvertes. Si l'obligation peut permettre d'avancer en ce sens (mandats de dépôt, identifiant requis par certains éditeurs, financeurs ou établissements), une appropriation par le chercheur reste la situation la plus propice à un engagement sur la durée, et un travail de formation est donc bienvenu.

Néanmoins, la manière la plus pertinente d'œuvrer à l'adoption d'identifiants par les chercheurs est d'œuvrer à la qualité de ces identifiants, à leur interopérabilité et aux services qui peuvent être proposés sur cette base, qui permettent au chercheur de voir la plus-value réelle de l'utilisation de tels identifiants.

Pour aller dans cette direction, l'intervention directe des professionnels de l'IST dans Orcid est possible, sinon dans la création elle-même de l'identifiant Orcid, du moins dans sa gestion, à travers la fonctionnalité de Trusted Organization<sup>225</sup> (Tiers de confiance) que le chercheur peut activer pour permettre à un système extérieur d'intervenir sur son compte Orcid (mises à jour, ajouts d'articles, de financements, etc.). Par ce biais, les professionnels de l'IST peuvent intervenir et soulager le chercheur d'une partie du travail de gestion des données associées à l'identifiant par le chercheur lui-même. Une synchronisation d'Orcid avec des systèmes locaux est proposée par Orcid<sup>226</sup>, et il est ainsi possible par exemple d'intégrer Orcid dans un système institutionnel de recensement des travaux des chercheurs, de manière à automatiser partiellement l'ajout de données et gagner en efficacité. Ainsi, un chercheur peut autoriser Datacite à accéder à son compte Orcid pour y verser ses jeux de données. L'Ifremer a aussi mis en place un système qui facilite la création d'un identifiant Orcid par le pré-remplissage de champs à l'aide d'informations issues de l'annuaire de l'Ifremer, et la mise à jour automatique de la bibliographie Orcid du chercheur à partir de l'archive ouverte Archimer<sup>227</sup> (fin 2020, 256 mises à jour automatiques étaient en place). C'est un exemple de la rationalisation que l'on peut opérer à travers les identifiants, permettant de rendre attractif leur usage, et aussi, dans le cas de l'Ifremer, de conforter l'obligation de dépôt dans Archimer à travers l'offre d'un service supplémentaire. C'est également un moyen, comme c'est le cas dans

<sup>224</sup> « Use of Author Identifier Services (Orcid, ResearcherID) and Academic Social Networks (Academia.Edu, Researchgate) by the Researchers of the University of Caen Normandy (France): A Case Study », Christophe Boudry et Manuel Durand-Barthez, *PLoS One* 15, 9, 2020, <https://doi.org/10.1371/journal.pone.0238583>. Consulté le 21 février 2021

<sup>225</sup> « Trusted Organizations », Orcid, <https://support.orcid.org/hc/en-us/articles/360006973893-Trusted-organizations>. Consulté le 3 février 2021

<sup>226</sup> « Workflows », Orcid, <https://info.orcid.org/documentation/workflows/>. Consulté le 3 février 2021

<sup>227</sup> Informations issues d'un retour d'expérience de l'Ifremer au consortium Orcid France en novembre 2020 (non public).

d'autres contextes, comme HAL (voir chapitre 4), d'introduire une curation professionnelle en interaction avec la gestion de leur identifiant par les chercheurs, même si la création en masse de comptes Orcid n'est pas vraiment possible.

#### *La gestion des référentiels de l'ESR : une mission à la fois nationale et locale*

Une bonne gestion des référentiels est nécessaire pour que les identifiants soient opérationnels. Cette gestion repose largement sur les professionnels de l'IST, aux échelles nationale et locale.

À l'échelle nationale, un problème est encore en attente de résolution pour les référentiels français : leur manque actuel de mutualisation entre eux et d'alignement systématique avec les référentiels internationaux. On a de nombreux référentiels, chacun créé et maintenu pour une fonction particulière, mais reproduisant partiellement le travail d'autres référentiels ayant une autre vocation.

Des entretiens avec des experts de ces questions nous ont conduit à penser que c'est le cas notamment pour les référentiels des structures et institutions de recherche, qui posent des problèmes de gestion particuliers, plus complexes que pour les référentiels auteurs, car ils sont organisés en plusieurs niveaux (labos, équipes, organismes, universités...), avec des liens historiques et structurels de succession, de fusion, etc. Il existe sur cette question, en France, plusieurs référentiels gérés en parallèle, pour ne pas dire parfois concurrents, dont les périmètres se superposent avec des rôles différents : le RNSR recense les structures au niveau national, mais il existe aussi des référentiels propres aux établissements – au CNRS ou à l'Inria par exemple –, avec lesquels le RNSR n'a pas de démarche de connexion. Au niveau supérieur, on trouve Sirene de l'Insee, qui recense seulement les structures ayant une identité juridique, avec hétérogénéité. Il y a ensuite Grid et ROR à l'international, mais ces référentiels s'adaptent mal au contexte français. Il y a enfin IdRef, Auréhal, ou encore Finess pour les établissements de santé.

Auréhal notamment est mal relié aux autres référentiels français, notamment à IdRef. La nature même du projet HAL, qui est de fonctionner par auto-archivage, implique de nombreux doublons car les chercheurs recréent des structures déjà présentes dans l'archive, ce qui implique un lourd travail de curation par les administrateurs de portail, et rend difficile l'alignement avec les autres référentiels. Plus profondément, ce type de problème est le reflet de la multiplicité des référentiels. Si l'ensemble des outils de l'ESR fonctionnaient sur la base d'un référentiel unique, par exemple IdRef, il serait possible d'imposer à l'utilisateur de choisir une institution parmi celles déjà référencées. Cela, notons-le, ne remet pas en cause en tant que tel l'auto-archivage. Bien sûr, cela nécessite que le référentiel en question soit exhaustif et précis, mais cela a d'autant plus de chances d'être le cas que les moyens sont fléchés vers une infrastructure ciblée et non plusieurs. Ces constats peuvent conduire à penser qu'il serait plus efficient qu'IdRef soit le référentiel nativement présent dans HAL, plutôt que de devoir lier entre eux *ex post* des référentiels de qualité variable.

Il serait techniquement possible d'optimiser l'architecture des référentiels de l'ESR, de mutualiser ces différentes bases locales et de tenter d'avoir un métaréférentiel solide, mais la multitude des acteurs aboutit à une certaine inertie. En conséquence, chacun poursuit la construction de son référentiel propre sans qu'une véritable mutualisation n'émerge jusque-là. Le problème est sans doute avant tout politique, alors même que parmi les référentiels cités, la plupart dépendent de la même gouvernance, le Mesri (c'est le cas de HAL, d'IdRef, du RNSR, du référentiel du CNRS, des référentiels internes aux

universités et EPST...), mais d'institutions différentes sous l'autorité de ce ministère et qui peinent à coopérer. La création en novembre 2020 d'un poste d'administrateur des données de l'ESR, occupé par Mme Isabelle Blanc<sup>228</sup>, doit permettre de remédier à certains de ces problèmes, en permettant d'avoir une vision et une stratégie globales, construites sur une priorisation des cas d'usages principaux et une coordination des acteurs impliqués.

Dans les établissements, un travail peut cependant être effectué pour améliorer les référentiels et les alignements entre eux. On verra dans le chapitre 5 comment certains établissements ont su créer des référentiels de qualité au sein d'un système d'information recherche. On peut signaler également ici un projet mené par l'archive Oatao à Toulouse en 2017 pour se connecter à IdRef<sup>229</sup>. Cela a permis un liage des auteurs de dépôts dans Oatao à l'autorité correspondante IdRef. Et, dans IdRef, les références bibliographiques et un lien vers le plein-texte des dépôts dans Oatao enrichissent les notices auteurs. Il est également possible de chercher dans Oatao par identifiant auteur IdRef, et de rebondir vers la page correspondante d'IdRef.

La France s'est également engagée sur la voie de l'alignement international des identifiants, avec la propositions du CoSO d'« adhérer au niveau national à Orcid, système d'identification unique des chercheurs qui permet de connaître plus simplement et sûrement les contributions scientifiques d'un chercheur »<sup>230</sup>. Couperin et l'Abes ont ainsi été mandatés par le Mesri pour piloter la naissance du consortium Orcid France. Cette communauté a été créée fin 2019 par 36 membres<sup>231</sup>. Cet investissement national dans Orcid est pensé en coordination avec IdRef : « L'Abes souhaite proposer Orcid en vitrine et IdRef en coulisse. IdRef demeure l'identifiant pivot (avec un spectre beaucoup plus large que les seuls chercheurs) sur lequel est conservée une maîtrise absolue. Avec l'identifiant Orcid, l'Abes entend donc promouvoir la visibilité internationale des chercheurs exerçant en France tout en restant garante de ces données d'autorité en s'appuyant sur IdRef, référentiel souverain pour l'ESR. »<sup>232</sup>

Un chantier est aussi en cours sur les identifiants d'organisations, qui à ce jour n'ont pas véritablement de standard international, malgré l'existence de ROR, Grid, Isni, Ringgold, mais « même si les standards ne sont pas totalement définis, il est important de rendre compatibles entre eux et avec le reste du monde les systèmes nationaux actuellement utilisés »<sup>233</sup>.

On voit donc que la gestion des référentiels doit être l'objet d'une synergie entre institutions internationales (agences de normalisation, organismes d'attribution, associations professionnelles et d'utilisateurs), nationales (l'Abes pour IdRef et Orcid, le CCSD pour HAL, le Mesri pour le RNSR, la BnF pour Isni et Vial) et locales pour construire les outils les plus pertinents possibles. Les prochains chapitres seront l'occasion de voir plus précisément comme cela peut se réaliser à travers la gestion des

<sup>228</sup> Profil LinkedIn d'Isabelle Blanc, Chief Data Officer au Mesri, <https://www.linkedin.com/in/isabelleblancatala/?originalSubdomain=fr>. Consulté le 19 février 2021

<sup>229</sup> « Déploiement d'Oatao dans IdRef : une nouvelle visibilité sur le web », Punktokomo, 2017, <https://punktokomo.abes.fr/2017/11/30/oatao-sappaue-sur-idref-pour-soffrir-une-nouvelle-visibilite-sur-le-web/>. Consulté le 3 février 2021

<sup>230</sup> « Des identifiants ouverts pour la science ouverte : note d'orientation », Comité pour la science ouverte, 2019, <https://www.ouvrirlascience.fr/des-identifiants-ouverts-pour-la-science-ouverte-note-dorientation/>. Consulté le 1<sup>er</sup> février 2021

<sup>231</sup> Isabelle Mauger Perez, « Consortium ORCID France : nouveau réseau, nouveaux correspondants », Arabesques n° 97, 2020, <https://publications-prairial.fr/arabesques/index.php?id=1772>. Consulté le 1<sup>er</sup> février 2021. Voir aussi une présentation d'Orcid France en vidéo : « Faciliter l'identification des personnes : la communauté Orcid France (Les Actus 2020 - partie 1) », Abes, <https://vimeo.com/415093274>. Consulté le 1<sup>er</sup> février 2021

<sup>232</sup> Ibid.

<sup>233</sup> « Des identifiants ouverts pour la science ouverte : note d'orientation », Comité pour la science ouverte, 2019, <https://www.ouvrirlascience.fr/des-identifiants-ouverts-pour-la-science-ouverte-note-dorientation/>. Consulté le 1<sup>er</sup> février 2021

archives ouvertes et par des procédures automatisées de curation des métadonnées.

### *Des référentiels disciplinaires en chantier*

Dans les établissements de l'ESR, les professionnels de l'IST trouvent également à employer leur expertise pour la mise en chantier d'enrichissements et d'alignements de référentiels très spécialisés. Cela peut notamment être l'objet d'appels à projets Collex, comme c'est le cas de Refdivinités, mené par la bibliothèque interuniversitaire de la Sorbonne (BIS). Ce projet a vocation « à faciliter l'indexation des documents et bases de données relatifs à l'Antiquité et à rapprocher IdRef, référentiel généraliste, de Pactols (Peuples, Anthroponymes, Chronologie, Toponymes, Œuvres, Lieux, Sujets), thésaurus spécialisé en archéologie »<sup>234</sup>. Plus précisément, les objectifs sont :

- de refondre et d'enrichir une partie de l'arborescence de Pactols (microthésaurus “Divinités” et “Héros” dans la partie “Anthroponymes”) ;
- de reprendre les notices d'autorités sur le même thème dans IdRef (...) ;
- d'aligner ces autorités entre elles, ainsi qu'avec Wikidata »<sup>235</sup>.

Un travail dans data.idref.fr a été réalisé pour explorer les données d'IdRef au moyen de Sparql, et dans le catalogue du réseau Frantiq (Fédération et ressources sur l'Antiquité) au moyen d'OpenTheso (un logiciel libre de gestion de thésaurus). Puis OpenRefine a été utilisé pour enrichir et aligner ces données de sources diverses, avant de créer de nouvelles notices dans Frantiq au moyen d'OpenTheso, et dans le Sudoc au moyen de WinIBW. Enfin, un chargement en masse des identifiants dans Wikidata par OpenRefine a été effectué<sup>236</sup>. Ce travail est progressivement complété, soutenu par la mise en place d'ateliers collaboratifs<sup>237</sup>.

Le projet rend ainsi possible un catalogue précis dans le Sudoc et dans le Catalogue collectif indexé du réseau Frantiq, et permet aussi de rebondir facilement vers Pactols et Wikidata, et à terme également vers le catalogue de Frantiq. Le traitement du corpus a par exemple introduit sur chaque notice « un qualificatif [qui] a systématiquement été utilisé pour préciser la qualité divine ou héroïque de l'entité décrite ainsi que l'aire civilisationnelle à laquelle elle appartient »<sup>238</sup>.

À la suite de ce projet, ArchéoRef Alignements (ArchéoAl) a pour but de produire des « enrichissements des autorités IdRef et Pactols de sites archéologiques et alignements de référentiels noms de lieux et géolocalisation »<sup>239</sup>.

On voit là un exemple de chantier mis en œuvre localement permettant d'enrichir des outils collectifs facilitant la recherche documentaire et le catalogue par entités. La force des bibliothèques académiques est de disposer de compétences propres à effectuer

<sup>234</sup> « Divinités et héros du monde méditerranéen antique : retour sur le projet RefDivinités », Punktokomo, 2020, <https://punktokomo.abes.fr/2020/11/19/divinites-et-heros-du-monde-mediterraneen-antique-retour-sur-le-projet-refdivinites/>. Consulté le 1<sup>er</sup> février 2021

<sup>235</sup> « Refdivinités », Collex-Persée, <https://www.collexpersee.eu/projet/interoperabilite-de-referentiels-sur-les-divinites-et-heros-du-monde-mediterraneen-antique/>. Consulté le 1<sup>er</sup> février 2021

<sup>236</sup> « Divinités et héros du monde méditerranéen antique : retour sur le projet RefDivinités », Punktokomo, 2020, <https://punktokomo.abes.fr/2020/11/19/divinites-et-heros-du-monde-mediterraneen-antique-retour-sur-le-projet-refdivinites/>. Consulté le 1<sup>er</sup> février 2021

<sup>237</sup> « Atelier d'alignement PACTOLS-Wikidata », Frantiq, 2021, <https://www.frantiq.fr/atelier-dalignement-pactols-wikidata/>. Consulté le 3 février 2021

<sup>238</sup> Ibid.

<sup>239</sup> « ArchéoRef Alignements (ArchéoAl) », Collex-Persée, <https://www.collexpersee.eu/projet/archeoref-alignements-archeoal/>. Consulté le 1<sup>er</sup> février 2021

ce travail dans un champ particulier, tandis qu'une agence nationale n'aurait pas les moyens humains ni les contacts parmi les chercheurs de ce champ pour mener l'ensemble de ces projets. De plus, la mise en œuvre locale de telles initiatives assure qu'elles répondent à un besoin réel, et permet en outre la mise en place de cocopérations entre chercheurs et professionnels de l'IST. D'autres exemples de projets Collex, tel Datapo (puis Datapoc 2.0) au MNHN<sup>240</sup> ou Mistara à la Bulac<sup>241</sup> en sont d'autres exemples.

## Bilan d'étape

Dans ce chapitre, nous avons montré l'importance des identifiants et des référentiels, qui servent de nœuds aux données liées du web sémantique. Mais nous avons également noté un certain nombre de défis, ayant trait par exemple à la diffusion encore modeste de certains identifiants, comme Orcid, ou au pilotage national des référentiels, avec des problèmes de qualité et de chevauchement entre référentiels, sur fond d'enjeux de gouvernance. Malgré tout, un référentiel comme IdRef montre qu'il est possible de construire un outil opérationnel et de susciter des convergences. Aux chapitres 4 et 5, nous verrons comment un travail sur les données, notamment par des procédures automatisées, permet de mieux répondre à certains de ces défis, même si les problèmes de gouvernance ne pourront pas trouver d'issue autre que politique.

---

<sup>240</sup> « Datapoc 2.0 », Collex-Persée, <https://www.collexpersee.eu/projet/datapoc-2-0/>. Consulté le 1<sup>er</sup> février 2021

<sup>241</sup> « Mistara », Collex-Persée, <https://www.collexpersee.eu/projet/mistara/>. Consulté le 1<sup>er</sup> février 2021

## CHAPITRE 4 : MÉTADONNÉES ET « INSIDE-OUT COLLECTION »

Les bibliothèques universitaires ne sont pas seulement des pourvoyeuses de ressources pour les chercheurs, elles accompagnent aussi, et de plus en plus dans le cadre de la science ouverte, la production scientifique de ces chercheurs et sa diffusion sous plusieurs formes : dépôts de publications en archives ouvertes, édition de revues ouvertes, exposition de données de la recherche... Lorcan Dempsey parle à ce sujet d'« *inside-out collection* »<sup>242</sup>. Au sein de cette dynamique, les bibliothécaires sont particulièrement bien placés pour apporter une aide significative aux chercheurs, notamment en termes de créations et traitement des métadonnées.

Sur ce plan, les exigences et recommandations du Plan S<sup>243</sup> (initiative de 2018 favorisant la mise à disposition en accès ouvert de toutes les productions scientifiques, sans embargo) pour les publications en accès ouvert et les archives ouvertes sont assez élevées en termes d'identifiants, de standards de métadonnées, de licence<sup>244</sup> : ainsi, l'usage des DOI est impératif, tout comme l'exposition d'un ensemble de métadonnées de haute qualité, dans un format interopérable, mentionnant des informations sur les financements, un statut OA lisible par les machines dans un format non propriétaire<sup>245</sup>. La conformité aux exigences d'Openaire en termes de métadonnées est recommandée<sup>246</sup>. Des identifiants pour les auteurs, institutions, financeurs le sont également fortement recommandés, tout comme l'inscription de la politique d'ouverture dans Sherpa-Roméo, la mise à disposition du plein-texte selon les standards Jats-XML (ou équivalent), ou encore la mise à disposition des données de citations selon les standards de l'Initiative for Open Citations (I4OC). Pour les archives ouvertes est aussi conseillée la mise à disposition des données via une API libre d'accès, ainsi que de liens entre les textes entiers et les métadonnées mises à disposition par des parties tierces comme PubMed, Scopus ou Crossref. À partir de ces exigences, on peut se demander dans quelle mesure elles sont satisfaites par les archives ouvertes en France et par les publications scientifiques gérées par les établissements.

### Quelle gestion des métadonnées en archives ouvertes ?

#### *HAL, un nœud dans l'économie des métadonnées de l'ESR français*

La « voie verte »<sup>247</sup> de l'Open Access consiste à exposer les productions scientifiques dans des archives ouvertes. En France, HAL, en tant qu'archive ouverte

<sup>242</sup> « Library Collections in the Life of the User: Two Directions », Dempsey, Lorcan. 2016, *Liber Quarterly* 26(4). <https://www.liberquarterly.eu/articles/10.18352/lq.10170/>. Consulté le 1<sup>er</sup> février 2021

<sup>243</sup> « Plan S Principles », cOAlition S, [https://www.coalition-s.org/plan\\_s\\_principles/](https://www.coalition-s.org/plan_s_principles/). Consulté le 3 février 2021

<sup>244</sup> « Part III: Technical Guidance and Requirements », cOAlition S, [https://www.coalition-s.org/technical-guidance\\_and\\_requirements/](https://www.coalition-s.org/technical-guidance_and_requirements/). Consulté le 3 février 2021

<sup>245</sup> Ibid.

<sup>246</sup> « Application Profile Overview », Openaire, [https://openaire-guidelines-for-literature-repository-managers.readthedocs.io/en/v4.0.0/application\\_profile.html#application-profile](https://openaire-guidelines-for-literature-repository-managers.readthedocs.io/en/v4.0.0/application_profile.html#application-profile). Consulté le 3 février 2021

<sup>247</sup> « La voie verte », Science ouverte France / Couperin, <https://scienceouverte.couperin.org/la-voie-verte-2/>. Consulté le 4 mars 2021

nationale, est le lieu privilégié de dépôt des productions de la recherche scientifique, autorisé par la loi pour une République numérique à être mises à disposition en accès ouvert après un embargo, pour celles financées sur fonds publics. Le CNRS a décidé en 2019 de n'alimenter les rapports d'activité de ses chercheurs que par des productions dans HAL, ce qui est une très forte incitation, voire une obligation de fait, à les y déposer. En mars 2021, l'ANR a annoncé que « toutes les publications issues des projets financés par l'ANR (a minima les versions acceptées pour publication) devront être déposées avec la licence CC-BY ou équivalente dans l'archive ouverte nationale HAL, immédiatement après publication conformément au Plan S »<sup>248</sup>.

Ces incitations font que l'archive ouverte HAL est centrale dans l'écosystème français de la science ouverte, des bases de données, des métadonnées et des référentiels ouverts. Elle est aussi un réservoir de métadonnées, même si le versement du plein-texte est fortement recommandé. Le versement des seules métadonnées des publications de la recherche française rendrait cependant déjà des services importants à de nombreux acteurs.

Des métadonnées obligatoires et optionnelles pour chaque type de documents sont spécifiées par HAL<sup>249</sup>. Il est possible de les récupérer automatiquement dans Crossref par le renseignement d'un DOI, voire à partir d'un PDF avec Grobid (Generation of Bibliographic Data).

### *Une intervention cruciale des professionnels de l'IST dans HAL*

Or les plateformes d'auto-archivage, et HAL notamment, posent des défis particuliers vis-à-vis des métadonnées puisque les dépôts se font, en principe, par les chercheurs eux-mêmes, ce qui peut occasionner des problèmes de qualité. Les organismes de recherche et leurs services documentaires fournissent en conséquence un travail important dans HAL pour améliorer le signalement, soit en faisant un travail de curation *a posteriori* soit parfois en se substituant au chercheur pour le dépôt.

Ainsi l'enquête Casuhal<sup>250</sup> parue en 2020 destinée à « évaluer les ressources humaines déployées en 2019 par les établissements dans la gestion quotidienne de leur portail et collections HAL » montre que le travail sur les métadonnées et référentiels est le premier poste en termes de ressources humaines, avec 21 % du temps de service consacré à HAL (soit 121 ETP, dont 72 de catégorie A, sur l'échantillon des répondants de 116 établissements de tailles variables<sup>251</sup>). Ce travail sur les métadonnées est pourtant également considéré comme une priorité pour les années à venir par 60 % des répondants (« Améliorer le référencement des publications/contrôle qualité des métadonnées »). De plus, parmi ces priorités on trouve également « Sensibiliser à la création d'identifiants pérennes de chercheurs (IdHAL, Orcid) » (74 %) et « Développer les actions autour des données de la recherche (formation, accompagnement, plan de gestion des données, liens données-publications, etc.) » (60 %). On voit donc qu'un travail déjà important est effectué sur les métadonnées de HAL, mais qu'il demeure seulement partiel puisque

<sup>248</sup> « Science ouverte : l'ANR prépare la mise en œuvre de la Stratégie de Non-Cession des Droits initiée par la cOAlition S », ANR, 2021, <https://anr.fr/fr/actualites-de-lanr/details/news/science-ouverte-lanr-prepare-la-mise-en-oeuvre-de-la-strategie-de-non-cession-des-droits-initiee-p/>. Consulté le 4 mars 2021

<sup>249</sup> « Déposer », HAL, <https://doc.archives-ouvertes.fr/deposer/>. Consulté le 4 février 2021

<sup>250</sup> « Enquête Adhérents 2020 – Synthèse des résultats », CasuHAL, 2020, <https://www.casuhal.org/2020/09/25/enquete-adherents-2020-synthese-des-resultats/>. Consulté le 4 février 2021

<sup>251</sup> Ce travail est le plus souvent, dans 69 % des cas, effectué par la bibliothèque de l'établissement, mais peut aussi être réalisé, conjointement ou indépendamment de la bibliothèque, par un autre service, notamment des laboratoires.

beaucoup d'établissements jugent nécessaire d'y consacrer du temps supplémentaire.

Pour Nicolas Alarcon, du SCD de l'université de La Réunion et du club des utilisateurs de HAL Casuhal, avec qui nous nous sommes entretenu<sup>252</sup>, la question des métadonnées dans HAL reflète une injonction contradictoire : on souhaite des métadonnées fiables et complètes, mais l'outil d'auto-archivage doit être manipulable par des chercheurs qui ne sont pas spécialistes de l'IST. On a donc d'un côté un objectif de simplification du dépôt dans HAL, mais pour garantir malgré tout une qualité minimale de métadonnées, il est nécessaire de déporter cette tâche vers des professionnels de l'IST. Comme le rappelle la feuille de route 2016-2020 du CCSD<sup>253</sup>, « faciliter le dépôt est une priorité, la qualité des métadonnées passant par d'autres flux ». Quels sont ces « autres flux » ? Ceux gérés par les professionnels de l'IST. Ainsi, l'objectif de simplification a amené à une diminution des champs obligatoires dans les dernières versions de HAL, suite à quoi une partie des métadonnées peuvent ne pas être renseignées par les déposants ; en conséquence, les professionnels de l'IST doivent ensuite compléter les métadonnées manquantes.

En outre, les déposants peuvent ne pas remplir correctement certains champs, par exemple mettre plusieurs informations dans le champ titre, que les professionnels doivent ensuite redistribuer dans les champs adéquats. Mais il n'existe pas de guide ni de charte référençant les bonnes pratiques, qui relèvent donc largement de l'implicite.

De plus, un déposant pouvant créer un auteur ou une affiliation dans HAL, cela implique un risque de multiplication de doublons. Le travail de curation de la part des gestionnaires de portail est donc important, notamment au niveau des affiliations des auteurs aux structures<sup>254</sup>.

Pour remédier à certains de ces problèmes, la stratégie suivie à l'université de La Réunion a été d'effectuer le dépôt à la place des chercheurs, une voie alternative à celle promue par le CCSD. Pour Nicolas Alarcon, on n'est toujours pas parvenu, au bout de vingt ans d'existence de HAL, à faire fonctionner de façon satisfaisante l'auto-archivage. Et la surcharge de travail chronique des chercheurs ne permet pas d'espérer une amélioration drastique de leur implication. Selon lui, seules des contraintes réglementaires comme celle du CNRS permettent d'améliorer vraiment le taux d'archivage. D'un autre côté, pour alléger le travail des chercheurs, il est possible de mettre en place un contrôle *a priori* et non *a posteriori* des dépôts. Casuhal estime ainsi que des publications sous licence Creative Commons doivent pouvoir être archivées dans HAL sans accord spécifique de l'auteur, de manière à ce que des professionnels de l'IST puissent récupérer de telles publications exposées par ailleurs pour les rapatrier dans l'archive ouverte. À l'université de La Réunion, cela se traduit par un taux de 70 % des dépôts réalisés par des professionnels de l'IST, même s'il est aussi proposé aux chercheurs de les former à l'auto-archivage. Notons que ce modèle n'est pas forcément transposable à tout établissement, notamment aux universités de recherche intensive. De plus, le Mesri et le CCSD restent attachés à l'auto-archivage dans HAL.

### ***Enrichir les métadonnées par des procédures automatiques***

À partir de là, l'enjeu selon Nicolas Alarcon est de parvenir à des procédures

<sup>252</sup> L'entretien s'est tenu le 30 octobre 2020. Certains points de vue sont les miens, d'éventuelles erreurs sont de ma responsabilité.

<sup>253</sup> « Feuille de route du CCSD 2016-2020 », CCSD, <https://www.ccsd.cnrs.fr/2016/06/feuille-de-route-du-ccsd-2016-2020/>. Consulté le 4 février 2021

<sup>254</sup> Voir à ce sujet le tweet de Nathalie Clot du 4 février 2021 à 11h04, au sujet des évolutions de HAL : <https://twitter.com/NaCl2/status/1357268907833389056?s=20>. Consulté le 5 février 2021



automatiques d'enrichissement des métadonnées des ressources dans HAL, pour permettre à la fois un dépôt facilité et une modération de l'inflation de la charge de travail pour les chercheurs comme les professionnels de l'IST.

Cette automatisation est rendue possible notamment par le développement d'outils de curation, ou de versement automatique. Par exemple, OcdHAL<sup>255</sup> est une application web développée par l'université de Grenoble, permettant de consulter le contenu d'une collection HAL, de compléter et corriger les métadonnées ou encore de repérer les publications pouvant faire l'objet d'un dépôt en texte intégral<sup>256</sup>. Un tel outil, s'appliquant à HAL mais développé à l'extérieur, demande cependant de la maintenance pour être compatible avec les évolutions implémentées par le CCSD. Une intégration plus avancée de cet outil (ou d'autres) serait donc bienvenue. On peut aussi signaler mHALoDOI, qui consiste en « une petite série de scripts (javascript) pour vérifier qu'une liste de DOI a bien des notices correspondantes déposées dans HAL avec la bonne affiliation » et « vérifier si des publications Orcid sont sur HAL ou trouver des DOI à partir d'une liste de titres ».

Il existe ainsi une série d'outils locaux permettant d'intervenir dans HAL<sup>257</sup>, développés dans les établissements, plus souvent par des informaticiens, mais en collaboration avec des professionnels de l'IST. Un travail est sans doute à faire pour élargir l'utilisation de ces outils, en leur accordant la publicité qu'ils méritent et de la formation<sup>258</sup>.

Cependant, ces outils fonctionnent souvent, comme on le voit avec mHALoDOI, à partir des DOI. Mais là aussi des obstacles surgissent : si pour les STM une très large majorité des publications récentes sont dotées d'un DOI, c'est beaucoup moins le cas en SHS et encore moins en sciences juridiques.

Un outil au spectre plus large, mais qui peut trouver une application pour le versement dans HAL, est Dissemin<sup>259</sup> : cet outil récolte des données de Crossref (pour les articles avec DOI), BASE (pour les autres articles, y compris preprints), Sherpa-Roméo (pour les politiques des éditeurs) et Zotero (pour les métadonnées associées aux DOI), et permet soit de trouver la page où un article est disponible en OA, soit de verser l'article dans un dépôt en accès ouvert (HAL, Zenodo, et peut-être prochainement dans ArXiv...) si ce n'est pas encore le cas. Le dépôt est simplifié puisque le chercheur n'a pas à resaisir les métadonnées, qui sont automatiquement récupérées. Dissemin utilise Orcid pour l'authentification et permet d'actualiser ses publications dans Dissemin à partir d'Orcid.

Un autre élément souligné par Nicolas Alarcon pour favoriser l'automatisation des dépôts est la mise à disposition des métadonnées des publications par les éditeurs avec le versement automatique de notices. C'est l'objectif par exemple du projet Collex Droit2HAL<sup>260</sup>, qui propose notamment « un traitement des métadonnées livrées par les éditions Dalloz visant à établir une conversion du XML Dalloz en XML-TEI HAL » et entend « permettre l'automatisation du dépôt des métadonnées collectées, et la mise en

<sup>255</sup> « OcdHAL », université Grenoble-Alpes, <https://ocdhal.univ-grenoble-alpes.fr/>. Consulté le 4 février 2021

<sup>256</sup> « OCdHAL : Tableau de bord HAL », université Rennes 1, <https://openaccess.univ-rennes1.fr/ocdhal-tableau-de-bord-hal>. Consulté le 4 février 2021

<sup>257</sup> Certains sont indiqués par le CCSD ici : « Outils et services développés localement pour améliorer ou faciliter l'utilisation de HAL », Wiki CCSD, [https://wiki.ccsd.cnrs.fr/wikis/hal/index.php/Outils\\_et\\_services\\_d%C3%A9velopp%C3%A9s\\_localement\\_pour\\_am%C3%A9liorer\\_ou\\_faciliter\\_l'utilisation\\_de\\_HAL](https://wiki.ccsd.cnrs.fr/wikis/hal/index.php/Outils_et_services_d%C3%A9velopp%C3%A9s_localement_pour_am%C3%A9liorer_ou_faciliter_l'utilisation_de_HAL). Consulté le 4 février 2021

<sup>258</sup> Un exemple de chaîne pour « Déposer et enrichir les publications d'une collection HAL avec overHAL, H2HAL et CrossHAL » a été présenté aux journées de l'Abes 2018, voir <https://fr.slideshare.net/abesweb/jabes-2018-posterriariennes1deposerenrichirpublicationshal>. Consulté le 4 février 2021

<sup>259</sup> « Bienvenue sur Dissemin », Dissemin, <https://dissem.in/>. Consulté le 4 février 2021

<sup>260</sup> « Droit2HAL », Collex-Persée, <https://www.collexpersee.eu/projet/droit2hal/>. Consulté le 4 février 2021

place d'un flux permettant le dépôt des notices HAL à chaque nouvelle publication de revue Dalloz ». L'intérêt de ce projet est de référencer la production en sciences juridiques dans HAL, son défaut est qu'il est restreint aux métadonnées et non aux textes intégraux. Pour autant, on sait les juristes frileux à rendre accessibles leurs productions en accès ouvert de façon générale, le simple référencement de leurs articles constitue donc déjà une avancée intéressante.

La licence nationale Elsevier Freedom Collection 2019-2022<sup>261</sup> prévoit également un versement des métadonnées<sup>262</sup> de certaines publications de cet éditeur dans HAL, via une API, mais avec des imprécisions sur le protocole et les conditions de réutilisation<sup>263</sup>. Cette décision a suscité des critiques du CCSD<sup>264</sup>, sur le protocole de versement des métadonnées dans HAL, sur les conditions de réutilisation des métadonnées et sur les conditions d'accès au manuscrit auteur accepté sur le site d'Elsevier.

### *Diversifier les métadonnées et améliorer les référentiels*

Un autre enjeu du signalement dans HAL concerne les schémas de métadonnées proposés. Des chantiers sur les champs de métadonnées sont envisagés ou en cours, pour diversifier les possibilités offertes et mieux s'adapter aux besoins. Par exemple, la possibilité de renseigner de façon plurielle le type de document serait très intéressant, mais n'est pas à ce jour disponible dans HAL, contrairement à d'autres archives. En effet, une intervention en colloque peut se transformer en un chapitre dans des actes de colloque, et il serait alors intéressant de pouvoir avoir plusieurs types de documents avec des champs spécifiques, ainsi que des documents principaux et annexes. Un groupe de travail est en cours de constitution pour répondre aux attentes des chercheurs qui ne trouvent pas forcément les types de documents répondant à leurs besoins (par exemple, des chroniques de jurisprudence pour les juristes, ou encore des comptes rendus d'ouvrages, des carnets de fouille, etc.). Un référentiel bien plus large que celui de HAL a été développé par Coar<sup>265</sup> et pourrait inspirer l'archive nationale française. Cette typologie est examinée dans le groupe de travail traitant actuellement ce sujet, selon le CCSD<sup>266</sup>.

Plus généralement, des chantiers sont envisagés ou en cours sur la qualité des référentiels ou l'alignement des référentiels HAL avec d'autres référentiels. AuréHAL dispose de trois référentiels modifiables par les utilisateurs pour les Auteurs, Structures et Revues. Les administrateurs de portail interviennent *a posteriori* sur ces référentiels pour faire du nettoyage et des fusions/suppressions (au moins une fois par an). Des

<sup>261</sup> « ScienceDirect (Freedom Collection) », Couperin, <https://www.couperin.org/negociations/liste-des-negociations/item/190-sciencedirect>. Consulté le 19 février 2021

<sup>262</sup> Lettre envoyée par la présidente du consortium Couperin à Elsevier le 11 avril 2019, <https://www.soundofscience.fr/wp-content/uploads/2019/04/2019-04-11-COUPERIN-Lettre-accord-Elsevier.pdf>. Consulté le 19 février 2021

<sup>263</sup> Comme le notait en 2020 Martin Clavey dans « The Sound Of Science » : « L'accord prévoit une mise en place automatique possible d'un accès au bout de 12 mois au "manuscrit auteur accepté" (MAA) ou postprint en streaming directement sur Sciencedirect, la plateforme d'Elsevier, ainsi qu'une notice HAL (l'archive ouverte du CNRS) qui pointe vers ce streaming. Puis, dans un deuxième temps et au bout de 24 mois, le fichier PDF de ce manuscrit se retrouverait directement sur la plateforme HAL ». Cependant, « cet accord permet donc à Elsevier de pousser les chercheurs français à ne surtout pas se préoccuper du dépôt de leurs articles en "green open access" en leur procurant un service qui le fait mais avec un embargo plus large que le permet la loi et en streaming et non avec le fichier PDF accessible directement », <https://www.soundofscience.fr/1754>. Consulté le 5 février 2021

<sup>264</sup> Voir le tweet d'Alain Marois du 6 juin 2019 : <https://twitter.com/amarois/status/1136683597413396480?s=20> (nous n'avons pas trouvé ce document en ligne hors de cette photographie sur Twitter) ; voir aussi « Lettre ouverte au consortium Couperin sur le renouvellement de l'abonnement à Elsevier », CAPSH, <https://blog.dissem.in/2019/lettre-ouverte-au-consortium-couperin-sur-le-renouvellement-de-labonnement>. Consultés le 5 février 2021

<sup>265</sup> « Controlled Vocabulary for Resource Type Genres (Version 2.0) », COAR, [http://vocabularies.coar-repositories.org/documentation/resource\\_types/](http://vocabularies.coar-repositories.org/documentation/resource_types/).

<sup>266</sup> Tweet du CCSD, du 5 février 2021 : [https://twitter.com/ccsd\\_fr/status/1357645619192561666?s=20](https://twitter.com/ccsd_fr/status/1357645619192561666?s=20). Consulté le 5 février 2021

référentiels Domaines, Projets ANR et Projets européens existent également, mais sont non modifiables.

Le référentiel Auteurs est en cours de refonte complète. En novembre 2020, il s'apprêtait à passer en bêta-test, pour une mise en production au printemps 2021. À l'heure actuelle, on observe une déconnexion entre forme auteur et IdHAL, ce à quoi cette évolution a pour objectif de remédier. Par ailleurs, les travaux dans HAL peuvent être versés dans Orcid depuis le 3 novembre 2020 via une manipulation simple<sup>267</sup>. L'inverse n'est pas actuellement proposé mais il s'agit d'un développement envisagé.

Un travail d'alignement IdHAL/IdRef a également été fait, mais sans interfaçage direct (voir au chapitre 3 les problèmes posés par cette situation). Il est en effet possible d'aligner les identifiants de deux manières :

- « si l'IdHAL est présent dans la notice IdRef, c'est qu'il a été ajouté à l'unité par un catalogueur dans la base de données IdRef ou que l'Abes a procédé à cet ajout dans le cadre d'un alignement automatisé par lot ;
- si l'IdRef est présent dans le compte HAL d'un auteur, c'est a priori le détenteur du compte qui l'a ajouté. »<sup>268</sup>

Une fois les deux identifiants liés, IdRef peut récupérer les publications dans HAL de l'auteur identifié par l'IdRef. De plus, l'Abes travaille à une « chaîne de traitements automatisée des références HAL où l'alignement des auteurs se fait directement à partir de la mention d'auteur (i.e. la chaîne de caractères “nom prénom”) présente dans les métadonnées de la publication HAL et les notices IdRef »<sup>269</sup>.

Un alignement du référentiel Structures avec le RNSR serait également bienvenu, même si ce référentiel national comporte lui-même des défauts. De même, le référentiel Revues de HAL gagnerait à être interfacé avec l'ISSN et le Sudoc, mais un travail préalable de nettoyage serait nécessaire (car on note que dans ce référentiel certains éléments ne sont pas normalisés – on peut trouver par exemple une URL ou une abréviation à la place du titre d'une revue).

Le référentiel Domaines est quant à lui assez pauvre, notamment côté SHS (seulement une vingtaine de catégories sont disponibles pour l'ensemble des SHS, contre des dizaines de domaines, avec une granularité très fine, pour la seule physique, en cohérence avec l'articulation de dépôt HAL / ArXiv).

Notons que selon un rapport de 2018, « les établissements qui souhaitent réaliser des analyses bibliométriques à partir des dépôts faits dans HAL, jugent la qualité des référentiels des auteurs et des structures de HAL insuffisante »<sup>270</sup>.

Enfin, les portails institutionnels peuvent ajouter des métadonnées propres sous la terminologie de « type locHAL », permettant l'ajout d'un champ spécifique (par exemple l'identification propre à une ancienne archive institutionnelle hors HAL qui rejoint HAL, par exemple l'Inrae quand ils ont rejoint HAL récemment pouvait ajouter un champ spécifique propre à leur archive). À noter que si ce champ n'est pas encore créé dans HAL, il s'agit d'une prestation payante du CCSD.

<sup>267</sup> « Compléter son Orcid iD avec ses dépôts HAL », CCSD, 2020, <https://www.ccsd.cnrs.fr/2020/11/compléter-son-orcid-id-avec-ses-depots-hal/>. Consulté le 5 février 2021

<sup>268</sup> « L'alignement des identifiants auteurs entre IdRef & HAL : un état des lieux », Punktokomo, 2020, <https://punktokomo.abes.fr/2020/10/02/lalignement-des-identifiants-auteurs-entre-idref-hal-un-etat-des-lieux/>. Consulté le 5 février 2021

<sup>269</sup> Ibid.

<sup>270</sup> « L'articulation des archives des établissements et de l'archive nationale pluridisciplinaire HAL », Cabinet Ourouk, 2018, <https://adbu.fr/competplug/uploads/2018/12/Etude-COPIST-4.pdf>. Consulté le 5 février 2021

Nous verrons plus loin qu'il est aussi possible d'intervenir de façon plus systématique sur une archive ouverte à travers des systèmes d'information recherche (chapitre 5).

## Quelles métadonnées pour les publications scientifiques en accès ouvert ?

Le chercheur qui souhaite diffuser sa production en accès libre, en évitant la « voie verte » (archives) au profit d'une forme éditorialisée mais alternative aux revues des grands éditeurs privé (qu'elle soient en accès fermé ou en « voie dorée » de l'accès ouvert, notamment avec APC) dispose d'une offre : des structures éditoriales offrent ces services relevant de la « voie diamant » de l'Open Access, souvent en interne des établissements de recherche ou universitaires, sous la forme de revue nativement numériques en libreaccès. Un certain nombre de ces publications scientifiques ne bénéficient pas, dans ce cadre, du soutien d'un éditeur professionnel (même si des éditeurs universitaires de haut niveau existent également). La gestion technique est alors souvent assurée par les bibliothécaires, notamment sur le traitement des métadonnées.

### *Prairial, un exemple de soutien aux productions éditoriales des établissements*

Prenons un exemple, local, donc partiel, mais représentatif, de ces projets éditoriaux institutionnels : le rapport « État des lieux et recommandations pour le soutien éditorial aux revues scientifiques du site Lyon-Saint-Étienne »<sup>271</sup> d'avril 2020 insiste sur les limites techniques que rencontrent les revues académiques étudiées, publiées par des structures internes : « Seule la revue *Focales*, créée en 2017, a pu bénéficier d'un site produit par le service informatique de l'université Jean-Monnet Saint-Étienne en utilisant le CMS Lodel développé par OpenEdition. Dans les autres cas, des CMS classiques ont été “bricolés” pour permettre la diffusion d'articles scientifiques : les métadonnées sont souvent absentes ou très pauvres, le contenu n'est pas structuré et normalisé, l'interopérabilité avec d'autres outils est rendue extrêmement complexe, la pérennité des contenus est douteuse, et la visibilité de la revue en pâtit. »

Ainsi, si les revues sur support numérique représentent pour les structures de recherche la possibilité de publier à coût plus réduit, une production trop artisanale ne permet pas d'en assurer la pérennité, la diffusion. Notamment, le référencement de ces revues n'est pas toujours optimal : « Plus d'une revue sur cinq n'est présente ni dans un catalogue de référence comme le Sudoc (...) ni sur un site dédié au référencement des revues comme Mir@bel. »

La raison en est la pauvreté des métadonnées renseignées (le rapport note l'« absence de métadonnées en Dublin Core et d'entrepôt OAI-PMH nécessaires au référencement par Isidore »<sup>272</sup>) et plus fondamentalement un manque de prise en compte des enjeux et exigences du signalement pour la visibilité des revues.

Face à ce constat, des initiatives de mutualisation peuvent être mises en place par

<sup>271</sup> « État des lieux et recommandations pour le soutien éditorial aux revues scientifiques du site Lyon-Saint-Étienne. Synthèse -avril 2020. » [Rapport Technique], Jean-Luc de Ochandiano, Alexandra Dugué, Laëtitia Le Couédic, Isabelle Bizos, Université Jean Moulin Lyon 3, Université Lumière Lyon2, MSH Lyon-Saint-Étienne, 2020 (hal-02732974), <https://hal-univ-lyon3.archives-ouvertes.fr/hal-02732974>. Consulté le 5 février 2021

<sup>272</sup> Ibid.

des professionnels de l'IST. Ainsi, Prairial<sup>273</sup>, née en 2017 et devenue en 2018 la première pépinière officielle d'OpenEdition, a ainsi permis d'aller vers une meilleure structuration du signalement des revues concernées, appuyée sur Lodel et Métopes. Ce dernier projet a vocation à soutenir l'activité éditoriale des établissements de l'ESR, par la normalisation des contenus et métadonnées des publications, notamment par l'utilisation des standards de la TEI<sup>274</sup>, tout en permettant une distribution multisupport et un archivage pérenne. En conséquence, Prairial a « largement consolidé son infrastructure technique et ses services (attribution de DOI aux articles, intégration de métadonnées auteur...) »<sup>275</sup>, mais la plateforme doit néanmoins aller plus loin encore pour répondre aux besoins, notamment en termes de diffusion à travers la mise en place d'un serveur OAI-PMH, d'un dépôt automatique dans HAL, et d'un référencement des revues dans Mir@bel, Isidore Openaire, Sudoc, Worldcat, DOAJ, Google Scholar, Sherpa/Roméo<sup>276</sup>... L'attribution de DOI notamment est cruciale pour la bonne identification des publications. On notera qu'en SHS, une large partie des publications n'en sont pas pourvues, ce qui a pour conséquence, entre autres, qu'elles ne sont pas prises en compte dans le baromètre de la science ouverte (voir chapitre suivant), donnant l'impression que le taux d'accès ouvert en SHS est bien plus faible qu'en réalité<sup>277</sup>.

### *OpenEdition, Persée, Cairn : des initiatives avancées dans la diffusion de la production scientifique*

Des initiatives plus avancées, mais qui touchent en conséquence des revues elles aussi plus établies, peuvent servir d'inspiration pour des initiatives plus locales, de type « pépinières » : dans le champ de l'open access, on peut souligner l'importance de Persée et OpenEdition. Cairn, initiative d'éditeurs privés, est également un pilier de la diffusion des publications de SHS en France.

**OpenEdition**<sup>278</sup> est une infrastructure de recherche qui anime un portail de publications en SHS créé par le Centre pour l'édition électronique ouverte, qui associe le CNRS, l'EHESS, l'université d'Aix-Marseille et l'université d'Avignon pour œuvrer à l'exposition de ressources en accès ouvert. Il diffuse les publications notamment de presses universitaires et sociétés savantes ayant un degré de maturité avancé : on dénombre (au 25 janvier 2021) 10 825 ebooks et 552 revues diffusées, ainsi qu'un calendrier scientifique sur Calenda et des carnets de recherche sur Hypothèses. Prairial étant une « pépinière » d'OpenEdition, les revues qu'elle prend en charge ont justement vocation à rejoindre à terme cette plateforme quand elles auront atteint un degré de maturité avancé.

OpenEdition est en mesure d'offrir des services de diffusion de métadonnées plus avancés. Ainsi, les données sont exposées en Dublin Core et en Mets dans un entrepôt OAI-PMH<sup>279</sup>, en Marc (Unimarc et Marc21)<sup>280</sup> pour les établissements ayant souscrit à la version Freemium (qui permet aussi de récupérer des fichiers PDF et ePub des

<sup>273</sup> « À propos », Prairial, <https://publications-prairial.fr/accueil/index.php?id=200>. Consulté le 5 février 2021

<sup>274</sup> « Text Encoding Initiative », TEI, <https://tei-c.org/>. Consulté le 5 février 2021

<sup>275</sup> « État des lieux et recommandations pour le soutien éditorial aux revues scientifiques du site Lyon-Saint-Étienne. Synthèse -avril 2020. », op. cit.

<sup>276</sup> « Offre de services », Prairial, <https://publications-prairial.fr/accueil/index.php?id=124>. Consulté le 5 février 2021

<sup>277</sup> Voir à ce sujet le tweet de Lionel Maurel le 25 janvier 2021 : <https://twitter.com/Calimaq/status/1353718343111487488?s=20>. Consulté le 5 février 2021

<sup>278</sup> Page d'accueil, OpenEdition, <https://www.openedition.org/>. Consulté le 5 février 2021

<sup>279</sup> « Dépôt OAI-PMH », OpenEdition, <https://www.openedition.org/8883>. Consulté le 5 février 2021

<sup>280</sup> « Notices au format Marc », OpenEdition, <https://www.openedition.org/8886>. Consulté le 5 février 2021

publications), et dans des fichiers KBart<sup>281</sup> pour l'exposition dans les outils de découverte ainsi que dans Bacon. Elle met aussi à disposition des équipes éditoriales un outil de vérification des métadonnées, CheckList<sup>282</sup>.

**Persée** est une initiative du Mesri datant de 2005, et visant à la mise à disposition en accès ouvert de collections numérisées de revues scientifiques. Elle développe un « accès libre aux contenus, développement open source, standards ouverts, interopérabilité et exploitation transversale des données »<sup>283</sup>. Le 13 octobre 2020, elle annonçait 858 457 documents diffusés. Persée assure un travail d'envergure sur les métadonnées et leur exposition, l'interopérabilité et la pérennité des données : structuration des contenus en XML-TEI, créations de métadonnées en Dublin Core, Mods, Marc-XML, autorités en Mads, alignement sur IdRef (et par ce biais sur d'autres référentiels, comme Isni, BnF...), attribution d'identifiants DOI et eISSN, exposition selon OAI-PMH et mise à disposition de fichiers KBart, données de citation sur Persée et dans Crossref (permettant des rebonds mais aussi des analyses bibliométriques), exposition en RDF dans le triplestore Data Persée<sup>284</sup>.

Évidemment, Persée traitant des collections rétrospectives, elle n'intervient pas en appui d'un travail éditorial en train de se faire, mais les développements qu'elle met en place sont néanmoins intéressants, et elle apporte un appui incontournable à la recherche en train de se faire grâce à ces services.

**Cairn** est une plateforme de publications scientifiques (revues et livres) créée en 2005 par les éditeurs Belin, De Boeck, La Découverte et Érès<sup>285</sup>, qui s'est ensuite élargie à de nombreux autres éditeurs en sciences humaines et sociales. Examinons cet exemple, qui s'il n'est pas intégralement en accès ouvert peut fournir des enseignements intéressants pour comprendre la chaîne de traitement des métadonnées par un éditeur.

Lors d'un entretien, Thomas Parisot<sup>286</sup>, co-directeur de Cairn, nous a présenté le circuit de création et d'exposition des métadonnées par la plateforme. Cairn reçoit des éditeurs un PDF non structuré, généré par InDesign le plus souvent ; l'enjeu est alors de pouvoir structurer le contenu de manière à le publier sous de multiples formats, y compris HTML. Au niveau des publications (revues ou ouvrages), à partir des PDF fournis par les éditeurs, Cairn produit une structuration XML normalisée pour l'ensemble des publications sur Cairn, format pivot permettant de créer des fichiers KBart ou Mets (descriptifs de collections), du Dublin Core pour le moissonnage en OAI-PMH, des notices Unimarc et Marc21. Cela permet ensuite d'alimenter en métadonnées des bases de connaissance (y compris Bacon) par KBart, les outils de découverte, catalogues, bases de données diverses où les productions sont signalées. En outre, des DOI sont attribués via Crossref au niveau des articles et chapitres de livre, ce qui permet de positionner les publications dans l'écosystème de métadonnées de Crossref. Signalons enfin qu'une intégration du référentiel Auteurs de Cairn et de celui d'IdRef a été lancée en janvier 2021, dans le cadre d'un partenariat entre l'Abes et Cairn : dorénavant, « les notices des nouveaux auteurs publiant dans les revues Cairn seront créées directement via IdRef »<sup>287</sup>,

<sup>281</sup> « Fichiers au format KBART », OpenEdition, <https://www.openedition.org/26973>. Consulté le 5 février 2021

<sup>282</sup> « Métadonnées et structuration des contenus : OpenEdition lance Checklist », OpenEdition, 2020, <https://leo.hypotheses.org/16687>. Consulté le 5 février 2021

<sup>283</sup> « Persée (portail) », Wikipédia, [https://fr.wikipedia.org/wiki/Pers%C3%A9e\\_\(portail\)](https://fr.wikipedia.org/wiki/Pers%C3%A9e_(portail)). Consulté le 5 février 2021

<sup>284</sup> « Publications scientifiques », Persée, <http://info.persée.fr/publications-scientifiques/>. Consulté le 5 février 2021

<sup>285</sup> « Cairn.info », Wikipédia, <https://fr.wikipedia.org/wiki/Cairn.info>. Consulté le 5 février 2021

<sup>286</sup> Entretien réalisé le 10 novembre 2020. Certains points de vue sont les miens, d'éventuelles erreurs sont ma responsabilité.

<sup>287</sup> « Cairn et l'Abes amorcent un partenariat autour d'IdRef », Fil'Abes, 2021, <https://fil.abes.fr/2021/01/15/cairn-et-labes-amorcent-un-partenariat-autour-didref/>. Consulté le 5 février 2021

afin d'éviter les doublons et de récupérer des métadonnées attachées à cet auteur et à ses publications.

Au sein de l'édition francophone en SHS, l'action de Cairn est cruciale, car elle permet aux 250 éditeurs participants, en vertu de sa taille critique, d'éviter de faire le travail de structuration des métadonnées, de devoir générer par eux-mêmes du KBart, de l'OAI, etc.<sup>288</sup> La mutualisation permet une montée en gamme importante en termes d'offre de services.

## Bilan d'étape

Les établissements de l'ESR peuvent être producteurs de métadonnées de deux façons : à travers les archives ouvertes et par des projets éditoriaux. Dans les deux cas, l'enjeu des métadonnées est crucial pour la visibilité du travail des chercheurs en accès ouvert, et requiert des compétences spécifiques en termes de curation des données, de politique d'identifiants, de protocoles d'échanges de données, de maîtrise d'outils et de connaissance du cadre politique et juridique. D'autres défis, déjà évoqués dans le chapitre 3, notamment d'administration des référentiels, sont également cruciaux pour ces questions. Nous y revenons dans le cinquième et dernier chapitre de ce mémoire.

---

<sup>288</sup> À partir des métadonnées, Cairn alimente un ensemble de bases de données (environ 150), ouvertes ou pas (Repec, Scopus...), ainsi que les index des outils découverte, où tout le contenu de Cairn est envoyé régulièrement (de façon hebdomadaire voire quotidienne). Cairn alimente aussi Crossref, ce qui permet d'alimenter de grandes bases par ce biais-là (Dimensions, The Lens par exemple). Suivant les cas, les métadonnées peuvent être moissonnées par OAI-PMH (Isidore), ou via des exports XML *ad hoc* selon des modalités et des formats propres aux cahiers des charges de telle ou telle base (Scopus par exemple).

## CHAPITRE 5 : LES TRAITEMENTS AUTOMATISÉS DE DONNÉES AU SERVICE DE LA QUALITÉ

Les professionnels de l'IST peuvent améliorer les métadonnées qu'ils reçoivent de diverses sources en leur appliquant des traitements automatiques. En raison des compétences spécifiques et encore assez rares requises pour effectuer ce type de tâches, ces traitements sont plutôt du ressort de grosses institutions, comme les agences bibliographiques. Dans un premier temps, nous nous pencherons donc sur des méthodes automatiques de curation des métadonnées qui rendent possibles leur traitement en masse, notamment à l'Abes. Cependant, les bibliothécaires dans les établissements peuvent aussi utiliser ces méthodes pour répondre à des besoins spécifiques. On le verra notamment sur la question des CRIS (Current Research Information Systems), qui permettent aux professionnels de l'IST de traiter les données *sur* la recherche (à distinguer des données *de* la recherche) de leur établissement, souvent à des fins de pilotage via notamment des indicateurs bibliométriques.

### La curation des métadonnées des éditeurs

Les métadonnées des éditeurs de publications scientifiques sont un premier espace d'application de méthodes automatiques de traitement des données par les professionnels de l'IST. On l'a déjà dit, les métadonnées des éditeurs souffrent parfois de faiblesses en termes de qualité, de normalisation des formats ou d'ouverture. Les bibliothécaires peuvent jouer un rôle sur ce plan, en adoptant ou développant des méthodes de traitement semi-automatique de ces données, mêlant automatisation et intervention humaine.

#### *Exemple 1 : le Hub de métadonnées de l'Abes et scienceplus.abes.fr*

En 2012 a été créé par l'Abes le Hub de métadonnées, aujourd'hui prolongé par la plateforme RDF scienceplus.abes.fr<sup>289</sup>. Le lancement de cette dernière est prévu en mars 2021 pour exposer des données RDF dans un triplestore, à un niveau de granularité fin (article, chapitre) qui constitue la grande spécificité de la base. Il s'agit donc d'un projet complémentaire du Sudoc, qui n'atteint pas une telle granularité (signalement au niveau du livre et non du chapitre, du numéro de revue et non de l'article), alors que les publications scientifiques peuvent être signalées plus finement avec profit. Il s'agit aussi d'une base provisoire : le projet de nouveau système de gestion des métadonnées que poursuit actuellement l'Abes<sup>290</sup> a vocation à élargir la couverture du Sudoc au-delà des entités qui y sont actuellement décrites. La base Scienceplus.abes.fr pourrait alors s'y intégrer. Sa philosophie n'est pas de devenir un index central de métadonnées, mais de traiter des corpus bien définis selon de hauts standards de qualité qui peuvent contribuer à cet index mondial en création. Le triplestore permet d'exposer ces données de façon souple de manière à laisser les utilisateurs se les approprier pour les besoins qui sont les

---

<sup>289</sup> Nous nous référons largement ici au rapport de 2013 sur le Hub de métadonnées [https://abes.fr/wp-content/uploads/2020/02/Hub\\_versionFinale\\_5juillet2013.pdf](https://abes.fr/wp-content/uploads/2020/02/Hub_versionFinale_5juillet2013.pdf), ainsi qu'à deux entretiens menés avec Yann Nicolas, expert métadonnées de l'Abes, le 11 janvier 2021 et le 17 février 2021.

<sup>290</sup> « Projet d'établissement 2018-2022 », Abes, 2018, <https://abes.fr/publications/publications-institutionnelles/projet-etablissement-2018-2022/>. Consulté le 28 février 2021



leurs. Si l'enjeu est bien sûr entre autres de mettre à disposition des outils de découverte des métadonnées de qualité de productions scientifiques, les usages potentiels sont multiples et à inventer par les professionnels de l'IST.

Le Hub de métadonnées, quant à lui, peut être défini comme un « atelier de retraitement en masse des données des éditeurs pour en tirer le maximum : en conserver toute la richesse, les enrichir encore et propager ces données de qualité dans différents environnements »<sup>291</sup>. Les métadonnées des éditeurs ne sont pas toujours de qualité optimale et demandent alors un important travail de curation. Dans le cadre de traitements de corpus massifs, comme ce fut par exemple le cas avec Istex<sup>292</sup>, la mise en place de procédures automatisées était nécessaire.

La stratégie du Hub était la suivante : au lieu de demander aux éditeurs de livrer leurs métadonnées dans un format particulier spécifique aux bibliothèques (Marc par exemple), au risque d'une perte de données importante ou d'une baisse de qualité, le Hub récupère les données des éditeurs en XML dans toute leur richesse, et les expose sans rien perdre (ou presque) en les convertissant en RDF. En effet, le modèle RDF est très souple et donc approprié à une telle conversion « universelle ». Sur cette base, on peut connecter les notices à des référentiels, les harmoniser selon des standards du web sémantique, compléter les données avec des identifiants, des URI, etc.<sup>293</sup>

La procédure de traitement au sein du Hub a donc consisté à transformer, au sein d'une base Oracle<sup>294</sup>, le fichier XML de départ en RDF-XML via XSLT, étape à laquelle était générée une URI pour chaque entité. Le fichier RDF-XML peut être transformé en triplets RDF automatiquement, et chargé dans une base RDF sous cette forme. Malgré cet aspect technique, ce workflow a été choisi entre autres parce qu'il ne nécessite pas de compétences informatiques poussées et permet une certaine autonomie technique des bibliothécaires.

Au sein de la base ainsi créée, en vertu de la souplesse de RDF, on peut enrichir les métadonnées avec les propriétés spécifiques dont on a besoin. Par exemple, pour associer aux articles des informations sur les auteurs avec la précision requise pour des publications scientifiques, en spécifiant qu'un auteur est le *premier* auteur ou le *dernier* auteur d'un article, ou qu'il écrit *en tant que* membre d'un institut de recherche, la propriété Dublin Core <dc:creator> est insuffisante, car elle se contente de spécifier un lien de création mais sans idée d'ordre ou d'affiliation. On a donc créé une propriété RDF « Authorship » *ad hoc* (dans l'espace de nom propre au projet), permettant d'intégrer ces informations plus précises. Une telle méthode évite de trafiquer un format standard.

Si on veut ajouter d'autres informations, comme des liens entre un auteur et IdRef ou Orcid, là encore le formalisme RDF est approprié : on peut utiliser l'ontologie OWL et la propriété <owl:sameas> pour affirmer, au sein d'un triplet RDF, que l'auteur est identique à l'entité correspondante à une certaine URI de IdRef ou Orcid<sup>295</sup>.

---

<sup>291</sup> « Le hub de métadonnées », Yann Olivier, Arabesques n° 83, 2016, <https://publications-prairial.fr/arabesques/index.php?id=526>. Consulté le 8 février 2021

<sup>292</sup> Page d'accueil, Istex, <https://www.istex.fr/>. Consulté le 20 février 2021

<sup>293</sup> « Autorités vs référentiels : 3 questions aux experts de l'Abes », Punktokomo, 2017, <https://punktokomo.abes.fr/2017/04/20/autorites-vs-referentiels-3-questions-aux-experts-de-labes/>. Consulté le 10 février 2017

<sup>294</sup> « Oracle Database », Oracle, <https://www.oracle.com/fr/database/>. Consulté le 20 février 2021

<sup>295</sup> Pour parvenir à ces alignements avec Orcid, on compare les DOI d'articles dans la base Istex et dans la base Orcid, on récupère les articles dont les DOI se trouvent dans les deux bases, puis en comparant les auteurs de ces articles, on peut repérer des similarités qui permettent de déduire que des auteurs sont identiques, et donc que l'Orcid spécifié dans la base Orcid peut être associé à l'auteur dans Istex.

Prenons l'exemple du traitement dans le Hub des revues et ebooks Springer<sup>296</sup>. Pour les revues, il s'agissait d'archives achetées en licence nationale de plus d'un millier de périodiques. Le Hub a permis de modéliser 1 965 510 articles à travers 71 millions de triplets RDF. Sur ces données ont été appliquées diverses méthodes de correction, par script Python ou OpenRefine, qui ont permis de détecter des anomalies (valeurs erronées, variantes de forme de titre non pertinentes, corrections d'ISSN électroniques...) ou des lacunes dans les collections livrées par l'éditeur.

Sur les ebooks Springer, un travail a été mené pour rattacher les contributeurs des ouvrages à des autorités IdRef. La stratégie suivie a été de retrouver les manifestations imprimées correspondant aux ebooks dans Worldcat à partir de l'ISBN renseigné par Springer, et à récupérer dans les notices de Worldcat ainsi identifiées les métadonnées sous forme de triplets RDF. On peut alors utiliser les triplets pointant vers les autorités Vial, pour ensuite les lier à IdRef. De plus, un webservice de Worldcat, Classify, a été exploité pour recueillir un indice Dewey à partir d'un ISBN, via une analyse des indices Dewey d'autres manifestations de la même œuvre dans Worldcat. Cette stratégie est représentative des protocoles mis en place dans ce type de procédure automatique : elle consiste à s'appuyer sur le modèle conceptuel Ifla-LRM pour aller récupérer des informations dans les notices d'une autre Manifestation de la même Œuvre, et les ajouter, quand c'est approprié, à la notice de la Manifestation dont on veut enrichir les métadonnées.

### *Exemple 2 : Métarevues, un outil pour générer l'historique de revues*

Présentons un autre exemple de procédure automatique développée dans le cadre du projet Istex, au moyen de l'outil Métarevues<sup>297</sup>, qui permet de générer à partir des données du Sudoc l'historique d'une revue à travers ses changements de nom, fusion, scission, etc. Dans le cadre de ce projet, les éditeurs sont amenés à fournir initialement une liste contractuelle qui, souvent, ne reflète pas exactement l'historique de ces revues, et est donc erronée. L'outil Métarevues permet de reconstituer la « métarevue » correspondant aux multiples instances de cette revue, en interrogeant de manière automatique les données du Sudoc, dans lequel les liens entre les différentes versions d'une revue sont en principe signalés. Ainsi, « l'outil remonte les informations des notices liées à partir des zones :

44X

- 440 Devient
- 441 Devient partiellement
- 442 Remplacé par
- 443 Remplacé partiellement par
- 444 Absorbé par
- 445 Absorbé partiellement par
- 446 Scindé en ... et en ...

---

<sup>296</sup> « Étude sur la faisabilité et le positionnement d'un hub de métadonnées Abes », Abes, 2013, pages 15 et suivantes, [https://abes.fr/wp-content/uploads/2020/02/Hub\\_versionFinale\\_5juillet2013.pdf](https://abes.fr/wp-content/uploads/2020/02/Hub_versionFinale_5juillet2013.pdf). Consulté le 8 février 2021

<sup>297</sup> Nous nous sommes entretenus à ce sujet avec Yann Nicolas lors d'un entretien le 11 janvier 2021. Voir aussi : « Métarevues : un outil dédié au traitement des périodiques », Punktokomo, 2014, <https://punktokomo.abes.fr/2014/06/26/metarevues-un-outil-dedie-au-traitement-des-periodiques/>. Consulté le 8 février 2021

- 447 Fusionne avec ... pour donner...

43X :

- 431 Suite partielle de
- 432 Remplace
- 433 Remplace partiellement
- 434 Absorbe
- 435 Absorbe partiellement
- 436 Fusion de ... et de ...
- 437 Séparé de
- 452 : a pour autre édition sur un support différent. »<sup>298</sup>

On obtient alors automatiquement l'intégralité des formes passées d'une revue. Cette procédure a permis d'intervenir dans le cours de la négociation Istex en pointant d'éventuelles erreurs, et de remodeler la liste contractuelle sur laquelle s'engageait l'éditeur.

L'outil Métarevues a ensuite pu être utilisé dans le cadre de Periscope, qui permet de comparer des collections de périodiques notamment dans le cadre des plans de conservation partagée<sup>299</sup>.

### *Exemple 3 : la gestion d'IdRef avec Paprika, une initiative collaborative soutenue par l'intelligence artificielle*

Si les exemples précédents concernent en premier lieu les agences bibliographiques, on peut donner un autre exemple de procédure semi-automatique impliquant à la fois l'Abes et les experts métadonnées dans les établissements de l'ESR : le contrôle qualité d'IdRef au moyen de Paprika.

L'application Paprika permet en effet un partage de la curation d'IdRef<sup>300</sup>. Elle est dédiée au contrôle qualité et à la création ou correction des liens entre les autorités IdRef et les notices du Sudoc, à travers une interface spécifique alternative à WinIBW (l'interface de catalogage du Sudoc)<sup>301</sup>.

Ce travail de gestion peut être amélioré par un outil de curation automatique, Qualinka, qui permet d'évaluer automatiquement la qualité de ces liens au moyen de l'intelligence artificielle<sup>302</sup>. Il fait des propositions que le catalogueur peut ensuite valider ou pas dans Paprika. Notons que Qualinka pourrait fonctionner hors de Paprika.

---

<sup>298</sup> Ibid.

<sup>299</sup> « Periscope », Sudoc, <https://periscope.sudoc.fr/>. Consulté le 8 février 2021

<sup>300</sup> « IdRef, Paprika and Qualinka. A toolbox for authority data quality and interoperability », Aline Le Provost et Yann Nicolas, 2020, <https://hal.archives-ouvertes.fr/hal-02563630>. Consulté le 8 février 2021

<sup>301</sup> Paprika a été mis en place dans le cadre d'une collaboration avec l'équipe GraphIK du Laboratoire d'informatique, de robotique et de microélectronique de Montpellier (Lirmm).

<sup>302</sup> « La curation, un enjeu pour la gestion des données numériques », Aline Le Provost, Arabesques n° 97, 2020, <https://publications-prairial.fr/arabesques/index.php?id=1793>. Consulté le 8 février 2021. Voir aussi la vidéo « Évolution des applications de signalement, pour faciliter les opérations de liage », de l'Abes, sur la curation de données avec Paprika, Qualinka, AlgoLiens, <https://vimeo.com/415122693>. Consultée le 8 février 2021

L'application s'appuie sur une modélisation du raisonnement tacite du bibliothécaire pour établir un diagnostic de coréférence ou de différence entre une autorité et un point d'accès d'une notice, avec un certain degré de confiance. L'ensemble du processus est synthétisé dans le schéma suivant<sup>303</sup> :

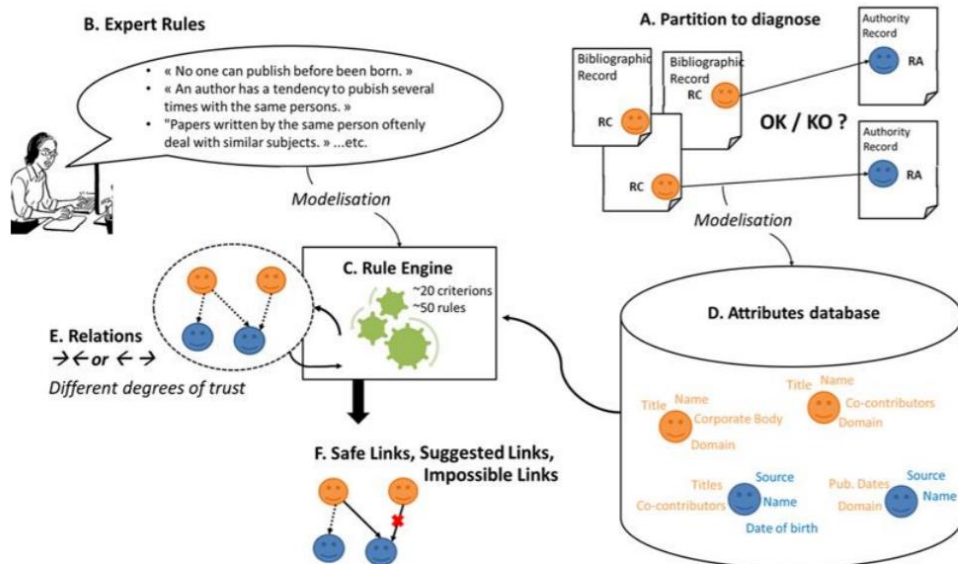


Fig. 6 : The trust values for co-reference / difference rules

Protocole de décision de Qualinka pour établir la coréférence entre une autorité et le point d'accès d'une notice<sup>304</sup>

On voit à travers ce projet comment une agence comme l'Abes et son réseau peuvent travailler de concert sur les données, avec la mise à disposition d'un outil d'intelligence artificielle accessible aux bibliothécaires pour effectuer un traitement semi-automatique.

## Le travail sur les données de l'ESR français et des établissements de l'ESR

Les professionnels de l'IST ont également un rôle à jouer dans la curation des données sur la recherche française, à l'échelle nationale comme dans les établissements, pour rendre plus visibles les acteurs et les productions de cette recherche.

### Exemple 1 : le référentiel Conditor

Le référentiel Conditor, développé par une équipe multipartenaire, a un statut incertain à l'heure de la publication de ce mémoire. Les développements au-delà de la phase projet n'ont pas à ce jour été rendus publics. Malgré tout, ce projet comporte de nombreux éléments qu'il nous semble intéressant de mentionner.

Conditor a vocation à répondre à la nécessité de rassembler des données sur la recherche française disséminées dans une multitude de bases (HAL, archives

<sup>303</sup> « IdRef, Paprika and Qualinka. A toolbox for authority data quality and interoperability », Aline Le Provost et Yann Nicolas, 2020, <https://hal.archives-ouvertes.fr/hal-02563630>. Consulté le 8 février 2021

<sup>304</sup> Crédit : Aline Le Provost, Yann Nicolas

institutionnelles, CRIS, etc.). Son but est donc de signaler « l'ensemble des productions scientifiques des établissements et laboratoires relevant de l'ESR (enseignement supérieur et la recherche) français et fournissant des métadonnées "qualifiées", via des API, à des applicatifs ESR identifiés »<sup>305</sup>. Pour cela, il vise à mutualiser les contributions des établissements en rassemblant et redistribuant des données (venant exclusivement de sources ouvertes), sur lesquelles il accomplit un travail de nettoyage et d'alignements en croisant des sources multiples, mais sans aucune saisie ni correction de métadonnées dans Conditor. Cette mutualisation permettrait d'éviter de nombreuses redondances dans les saisies de métadonnées.

Dans le livre blanc *Pour une meilleure visibilité de la recherche française*<sup>306</sup>, Antoine Blanchard et Elifsu Sabuncu présentaient en 2015 le schéma suivant, lui-même issu des animateurs du projet, pour résumer l'intention de Conditor :

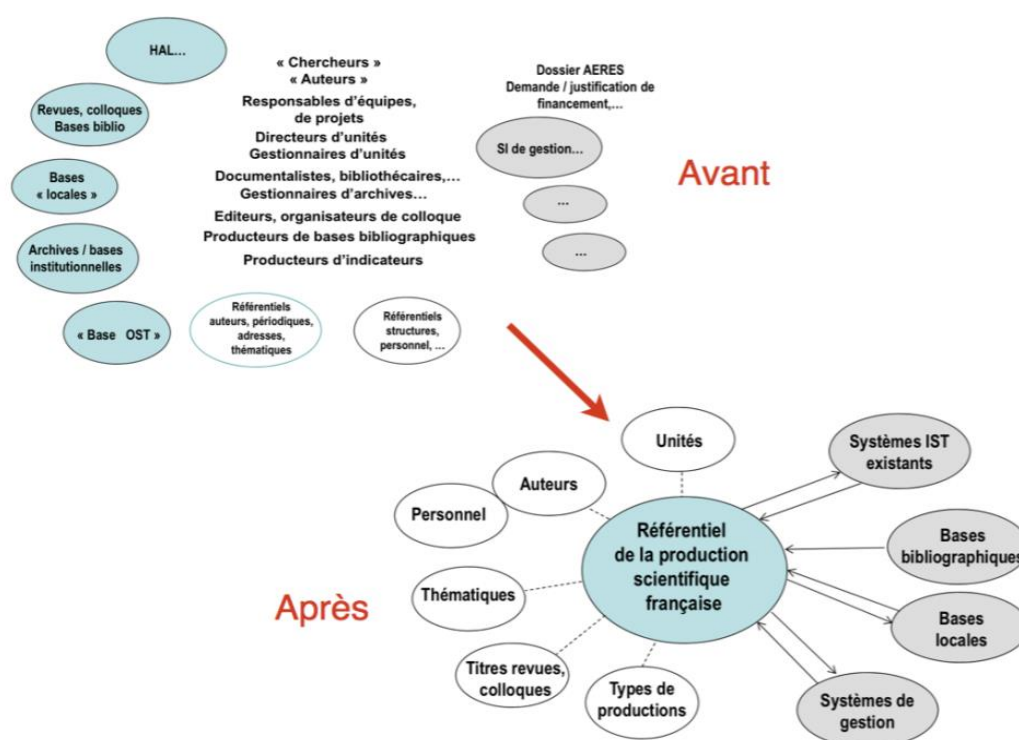


Schéma du projet Conditor dans l'écosystème de la production scientifique française<sup>307</sup>

Le projet Conditor est pensé pour recueillir, traiter et redistribuer les métadonnées en articulant :

- « un "pot" commun de métadonnées alimenté par des réservoirs sources ESR ou internationaux et aligné sur le Répertoire national des structures de recherche (RNSR) et des référentiels choisis collégalement,
- un "outillage" adapté pour le construire (collecte, reformatage, détection des doublons, enrichissement via des référentiels, etc.),
- un réseau de professionnels de l'information pour la qualification des données

<sup>305</sup> « Ce que n'est pas Conditor », Wiki Conditor,

[https://wiki.conditor.fr/conditor/index.php/Conditor\\_en\\_bref#Ce\\_que\\_n.E2.80.99est\\_pas\\_Conditor](https://wiki.conditor.fr/conditor/index.php/Conditor_en_bref#Ce_que_n.E2.80.99est_pas_Conditor). Consulté le 8 février 2021

<sup>306</sup> « Pour une meilleure visibilité de la recherche française », Antoine Blanchard et Elifsu Sabuncu, 2015, <https://hal.archives-ouvertes.fr/hal-01251541/document/>. Consulté le 8 février 2021

<sup>307</sup> Source : Annie Coret et Raymond Berard, journées Renatis, Cachan, 4 juillet 2012.

(validations, alertes vers les sources, etc.),

- des API permettant aux applicatifs de l'ESR de récupérer des métadonnées. »<sup>308</sup>

La structure du projet initial peut ainsi être synthétisée par le schéma suivant<sup>309</sup> :

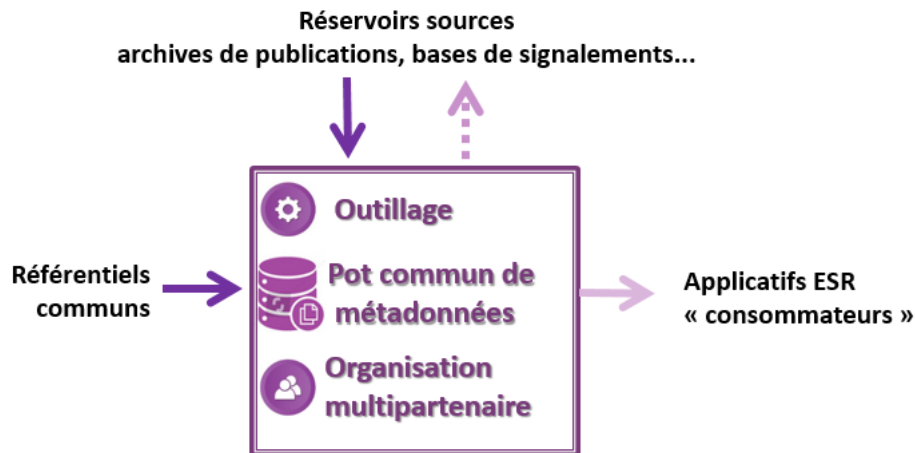


Schéma synthétisant le workflow de Conditor

Un bilan d'étape a été publié en juillet 2020<sup>310</sup>. Un travail de collection et de traitement a été effectué au cours de la phase projet sur quatre corpus sources (HAL, CrossRef, PubMed et les thèses du Sudoc). Le traitement semi-automatique des données comprend un reformatage dans un format pivot qui est « une extension de la TEI HAL afin de faciliter l'articulation avec l'archive nationale : il comprend des métadonnées bibliographiques mais aussi de gestion pour assurer la traçabilité de toute métadonnée »<sup>311</sup>. Ce travail, dans la mesure où aucun format d'entrée n'est imposé, nécessite un travail important en amont pour la création de feuilles de style XSLT, pour chaque corpus. Conditor permet ensuite la détection de doublons certains et incertains, et propose dans une interface nommée Cornelius un espace de validation des doublons incertains dans lequel interviendraient les professionnels de l'IST dans les établissements. Le projet prévoit également des modules d'alignements certains et incertains avec le RNSR, avec également une interface de validation par le réseau métier.

Au cours de la phase projet, 1 430 256 signalements ont été intégrés, décrivant au moins 1 108 307 productions distinctes (il demeure donc 145 011 signalements ayant un ou plusieurs doublons incertains).

Ce projet ambitieux se heurte néanmoins à des difficultés de mise en place. En novembre 2020, le projet s'est donc repositionné en priorité sur le seul workflow d'alimentation de Conditor vers HAL, avec un financement ad hoc<sup>312</sup>. En revanche, le réseau métier spécifique est abandonné, et il a été jugé préférable de s'appuyer sur les

<sup>308</sup> « Ce que n'est pas Conditor », Op. Cit.

<sup>309</sup> Ibid.

<sup>310</sup> « Conditor – bilan », Équipe Conditor, 2020, <https://www.ouvrirlascience.fr/conditor-3/>. Consulté le 8 février 2021

<sup>311</sup> Ibid.

<sup>312</sup> Conseil scientifique de l'Abes, 13 novembre 2020.

administrateurs de HAL pour effectuer les dédoublonnages manuels.

Dans la réunion de fin de phase projet, Bénédicte Kuntziger, du CCSD, estimait que Conditor permettrait :

- « – d’alimenter une partie référence bibliographique de Hal,
- de détecter des doublons intra Hal certains et incertains,
- d’enrichir des notices Hal »<sup>313</sup>.

L’intérêt de Conditor pourrait néanmoins s’étendre au-delà de HAL, par exemple aux archives institutionnelles, telle UnivOak de l’université de Strasbourg. Ainsi, un outil de type Conditor pourrait permettre, pour cette archive institutionnelle, selon Adeline Régé, responsable du Pôle Appui à la diffusion de la recherche :

- « – une fiabilisation des données en bénéficiant des enrichissements de Conditor,
- l’exhaustivité du fait du repérage dans d’autres sources de signalements (PubMed par exemple),
- une simplification du dépôt pour le rétrospectif : si les métadonnées sont fournies par Conditor, les chercheurs seront sollicités pour avoir le texte intégral »<sup>314</sup>.

Des outils comme Conditor pourraient en outre contribuer à des projets de type CRIS, comme à l’EHESS, ou Caplab<sup>315</sup>, ou encore les baromètres de la science ouverte. On voit donc que ce type d’outil, pourvu qu’il soit approprié par les professionnels de l’IST dans les établissements, peut enclencher une dynamique où ces professionnels sont à la fois utilisateurs, ce qui leur permet de gagner en efficacité sur certaines tâches, et potentiellement contributeurs puisqu’ils vont apporter des données et participer à leur curation, voire à l’élaboration de nouvelles fonctionnalités.

### **Exemple 2 : les baromètres de la science ouverte**

Le Baromètre de la science ouverte (BSO), publié pour la première fois en 2019<sup>316</sup>, et récemment rendu public pour 2020<sup>317</sup>, permet d’évaluer la part de la production scientifique française accessible en accès ouvert, selon les disciplines, les éditeurs, et le lieu d’hébergement (chez l’éditeur ou sur une archive ouverte comme HAL)<sup>318</sup>. Dans la dernière publication, traitant de l’année 2020, « 56 % des 156 000 publications scientifiques françaises publiées en 2019 sont en accès ouvert en décembre 2020 », tandis que « le taux observé en décembre 2019, relatif aux publications réalisées en 2018, n’était que de 49 % ».

Je me suis entretenu avec Emmanuel Weisenburger et Éric Jeangirard<sup>319</sup>, du service d’aide à la décision du Mesri, à propos de ce projet et de ScanR, une autre initiative de

<sup>313</sup> « Compte-rendu de la réunion de fin d’étape de la phase projet tenue le 6 février 2020 », Wiki Conditor, [https://wiki.conditor.fr/conditor/images/3/3d/Conditor\\_CR\\_20200206\\_bis.doc.pdf](https://wiki.conditor.fr/conditor/images/3/3d/Conditor_CR_20200206_bis.doc.pdf). Consulté le 8 février 2021

<sup>314</sup> Ibid.

<sup>315</sup> Voir les retours d’expériences correspondants, pour l’EHESS et Caplab : « Compte-rendu de la réunion de fin d’étape de la phase projet tenue le 6 février 2020 », Ibid.

<sup>316</sup> « Baromètre de la science ouverte : 41 % des publications scientifiques françaises sont en accès ouvert », Mesri, 2019, <https://www.enseignementsup-recherche.gouv.fr/cid146157/barometre-de-la-science-ouverte-41-des-publications-scientifiques-francaises-sont-en-acces-ouvert.html>. Consulté le 9 février 2021

<sup>317</sup> « Baromètre français de la Science Ouverte 2020. Note Flash n°1 », Mesri, 2021, <https://www.enseignementsup-recherche.gouv.fr/cid156502/barometre-francais-de-la-science-ouverte-2020.html>. Consulté le 9 février 2021

<sup>318</sup> « Le baromètre de la science ouverte », Comité pour la science ouverte, 2020, <https://www.ouvrirelascience.fr/le-barometre-de-la-science-ouverte/>. Consulté le 9 février 2021

<sup>319</sup> Le 3 décembre 2020. Certains points de vue sont les miens, d’éventuelles erreurs sont ma responsabilité.

recueil massif et d'exploitation de données sur l'ESR. La méthodologie du Baromètre de la science ouverte<sup>320</sup> est elle-même ouverte, comme les sources utilisées et les données de sortie : l'initiative est fondée intégralement sur l'open data. Des outils d'intelligence artificielle ont été développés pour mener à bien le projet. La procédure consiste dans un premier temps à identifier les publications dotées d'un DOI ayant une affiliation française. Parmi les 4,5 millions d'articles publiés chaque année avec un DOI, il s'est agi d'abord de vérifier si ces articles étaient présents dans HAL, et à défaut à lancer un parseur sur les pages de redirection du DOI pour tenter d'identifier un élément dénotant une affiliation française (à partir d'une liste d'expressions comme « France » et les principales villes françaises). Le fait de récupérer l'information de façon automatique directement sur les sites web permet de contourner l'absence de mise à disposition ouverte, via des API, de ces informations par les éditeurs.

Dans un second temps a été effectué un travail sur les données, permettant d'identifier les auteurs, les affiliations et les champs disciplinaires des publications, en enrichissant les métadonnées par les identifiants Orcid, IdRef, HAL pour les auteurs, et Grid, RNSR et Sirene pour les affiliations. Pour identifier les champs disciplinaires, une méthode d'intelligence artificielle fondée sur un apprentissage supervisé (avec un entraînement sur des données de Pascal et Francis) a été utilisée, permettant de déterminer le champ disciplinaire à partir des mots du titre, en suivant l'approche FastText<sup>321</sup>. Enfin, le statut Open Access ainsi que des données sur la licence et le lieu d'hébergement de la ressource ont été collectées en utilisant des données de HAL et Unpaywall<sup>322</sup>.

Un projet de Baromètre Santé de la science ouverte est actuellement en développement, déclinaison plus approfondie sur les ressources dans le secteur de la santé (en lien avec la crise sanitaire en cours), qui intégrerait un volet avec des informations sur les APC tirées du DOAJ, et un volet sur les essais cliniques, pour observer le pourcentage de déclaration publique des essais cliniques et mesurer le taux de transparence, outil qui pourrait permettre d'éviter un biais au résultat positif (en effet, les essais déclarés publiquement sont surtout des essais réussis, ce qui peut donner l'impression que la proportion d'essais fructueux est plus importante qu'en réalité).

Lié au BSO et développé par la même équipe du Mesri, ScanR est un moteur de recherche permettant d'explorer le paysage de la recherche française, qui peut faire ce que font des bases payantes comme Scopus ou WoS, mais avec des données ouvertes et redistribuant des données ouvertes, et de façon plus complète sur le périmètre France. Mi-janvier 2021, il recensait plus de 42 000 entités de recherche, plus de 500 000 fiches auteurs, près de 100 000 financements et plus de 500 000 productions scientifiques (publications, thèses et brevets)<sup>323</sup>, ainsi que des projets de recherche, qu'il relie entre eux.

ScanR s'appuie notamment sur IdRef et le RNSR. Il inclut des données du BSO mais prend aussi en compte HAL pour les publications sans DOI, publiées depuis 2013, ainsi que les thèses de Theses.fr depuis 1990 et les monographies indexées dans le Sudoc depuis 2013. Il pourrait, à terme, généraliser cette approche en déterminant de façon automatique, en crawlant les pages des DOI, toutes les affiliations, y compris non françaises.

---

<sup>320</sup> « Monitoring Open Access at a national level: French case study », Éric Jeangirard, 2019, ELPUB 2019, 23rd edition of the International Conference on Electronic Publishing, juin 2019, Marseille, France, <https://hal.archives-ouvertes.fr/hal-02141819>. Consulté le 9 février 2021

<sup>321</sup> Page d'accueil, FastText, <https://fasttext.cc/>. Consulté le 9 février 2021

<sup>322</sup> Page d'accueil, Unpaywall, <https://unpaywall.org/>. Consulté le 9 février 2021

<sup>323</sup> Page d'accueil, ScanR, <https://scanr.enseignementsup-recherche.gouv.fr/>. Consulté le 9 février 2021



On le voit de nouveau sur ces deux projets, des méthodes automatiques supervisées peuvent être très utiles pour traiter des métadonnées en masse et obtenir des résultats très pertinents, permettant en l'occurrence de valoriser la science ouverte. Mais selon Emmanuel Weisenburger et Éric Jeangirard, les données de départ posent des problèmes de qualité, et un engagement des professionnels de l'IST dans les établissements serait utile pour rendre plus facile la construction d'outils comme le BSO ou ScanR.

Un projet comme ScanR, se positionnant en bout de chaîne, se heurte aux problèmes de qualité des référentiels qui les alimentent. Ces problèmes relèvent souvent de la gouvernance au niveau national, mais d'autres aspects relèvent plus spécifiquement de l'action des acteurs locaux. On peut par exemple déplorer la faiblesse de l'exposition ouverte des annuaires de nombreuses institutions de recherche. Ainsi, « ScanR n'utilise pas les annuaires car ils ne sont pas rendus publics sous une forme exploitable par les institutions d'enseignement supérieur et de recherche »<sup>324</sup>. Par ailleurs, les erreurs dans ScanR peuvent également provenir de qualité des sources et des référentiels, avec leurs erreurs, leurs imprécisions, leurs omissions...

Les professionnels de l'IST pourraient donc jouer un rôle dans leurs établissements pour rendre ces sources d'informations d'une part de meilleure qualité, d'autre part plus ouvertes. Cela nécessite cependant de développer des connaissances et des formations pour que davantage de professionnels de l'IST acquièrent la culture non seulement technique, mais aussi scientifique et institutionnelle nécessaire à une gestion rigoureuse de référentiels, notamment Auteurs ou Structures.

Au-delà de l'amélioration à la source, les professionnels de l'IST et les chercheurs avec lesquels ils travaillent sont aussi des utilisateurs potentiels de ScanR. Donnons des exemples d'utilisations précises : ScanR peut être utilisé via son API pour la détection de conflits d'intérêts dans la constitution d'un jury par exemple, puisqu'il permet de repérer si des chercheurs ont travaillé ensemble. On peut aussi imaginer utiliser ScanR pour effectuer des recommandations sur des appels à projets européens, sur lesquels les laboratoires français candidatent moins que d'autres pays. L'équipe de ScanR aimerait donc mettre au point un outil pour identifier le laboratoire adéquat pour répondre à tel ou tel appel à projets, ce qui nécessite d'avoir des informations riches sur l'activité intellectuelle d'un laboratoire, à partir des mots-clés des publications par exemple, puis de lancer une opération d'intelligence artificielle permettant de lier ces spécialités aux appels à projets.

Suite à la publication du BSO national, plusieurs universités ont lancé des BSO locaux en reprenant les codes et méthodes mis à disposition par le Mesri. Ainsi, un outil lancé par un acteur national peut engendrer des déclinaisons locales, pourvu qu'il soit ouvert d'une part, et que des professionnels de l'IST se le réapproprient d'autre part, ce qui peut impliquer des choix en termes de ressources humaines, de budget et de formation. Une adaptation du code du BSO pour les universités a été produite par l'université de Lorraine<sup>325</sup>, mise à disposition et réutilisée par d'autres universités, telles celles de Saclay<sup>326</sup>, Versailles-Saint-Quentin-en-Yvelines (UVSQ)<sup>327</sup>, Évry<sup>328</sup> ou encore

---

<sup>324</sup> « Foire Aux Questions », ScanR, <https://scanr.enseignementsup-recherche.gouv.fr/faq>. Consulté le 9 février 2021

<sup>325</sup> « Baromètre lorrain de la science ouverte », université de Lorraine, <http://scienceouverte.univ-lorraine.fr/barometre-lorrain-de-la-science-ouverte/>. Consulté le 9 février 2021

<sup>326</sup> « Le baromètre de la science ouverte de l'université Paris-Saclay », université Paris-Saclay, <https://www.universite-paris-saclay.fr/barometre-science-ouverte>. Consulté le 9 février 2021

<sup>327</sup> « Baromètre Science Ouverte UVSQ », université Versailles-Saint-Quentin, <https://www.uvsq.fr/barometre-science-ouverte>. Consulté le 9 février 2021

<sup>328</sup> « Baromètre de la science ouverte d'Évry », université d'Évry, <https://www.biblio.univ-evry.fr/index.php/recherche-et-enseignement/science-ouverte/barometre-de-la-science-ouverte-evry/>. Consulté le 9 février 2021

Strasbourg<sup>329</sup>. L'UVSQ a de nouveau approfondi le travail en étendant la couverture aux publications sur HAL sans DOI, en obtenant des informations sur les APC, et en quantifiant les publications en Open Access selon le type (Diamond, Gold, Green...)<sup>330</sup>.

### *Exemple 3 : les CRIS ou systèmes d'information recherche*

Nous avons déjà constaté une certaine dispersion des données et référentiels existants entre une multitude de bases différentes. Cela ne facilite pas la saisie des données par les chercheurs, qui en conséquence de cet éparpillement doivent saisir les mêmes informations dans plusieurs outils. Dans leur *Livre blanc pour une meilleure visibilité de la recherche française*<sup>331</sup>, en 2015, Antoine Blanchard et Elifsu Sabuncu estimaient que les saisies de données devraient répondre au principe : « *input once, output many* ». Mais ils constataient qu'en réalité, « chaque étape est gérée dans un système d'information spécifique : les publications sont gérées dans des outils de gestion bibliographique et stockées dans des archives ouvertes, les données de la recherche alimentent des bases de données spécialisées ou prennent la poussière sur un disque dur, chaque agence de financement a sa propre plateforme de soumission des projets de recherche, les sites web de laboratoire ne se mettent pas à jour automatiquement ». En conséquence, soit les données doivent être renseignées plusieurs fois, soit elles ne seront pas saisies systématiquement, au détriment de l'exhaustivité et de la qualité de l'enregistrement en question.

En outre, Emmanuel Weisenburger et Éric Jeangirard<sup>332</sup>, du Mesri, notent une faiblesse dans la mise à disposition des métadonnées d'affiliation (liens entre les auteurs d'un article et leurs institutions de rattachement) de façon ouverte, car les bibliothèques universitaires réalisent un travail sur ces données, mais qui est ensuite mis au profit d'acteurs commerciaux (notamment Clarivate-Web of Science, afin de favoriser le classement de leur établissement dans les classements internationaux), et pas forcément vers des outils ouverts comme HAL, ce qui les oblige à refaire le même travail dans chaque outil. Un système distribuant l'information selon plusieurs canaux de sortie pourrait permettre de résoudre ce type de problèmes.

Une partie de la solution à ces défis pourrait reposer sur les systèmes d'information recherche, ou CRIS (Current Research Information System), qui permettent d'agréger et de traiter les informations sur la recherche, les productions scientifiques et les chercheurs d'un établissement. Ces informations proviennent de sources diverses, dont le SCD, mais aussi la direction de la recherche qui anime le suivi bibliométrique, les services Ressources humaines qui disposent d'informations sur les chercheurs, les services financiers qui détiennent des données sur les financements des projets... Les initiatives de type BSO locaux et la gestion de portails HAL peuvent être intégrées dans ce type d'initiatives qui permettent une gestion efficace de métadonnées *de* et *sur* la recherche dans un établissement à des fins internes de pilotage et de bibliométrie notamment, ou externes de visibilité de la production scientifique.

---

<sup>329</sup> « Le baromètre de la science ouverte de l'Unistra », université de Strasbourg, <https://scienceouverte.unistra.fr/strategie/le-barometre-science-ouverte/>. Consulté le 9 février 2021

<sup>330</sup> Voir le tweet de Maxence Larrieu du 5 février 2021 : <https://twitter.com/ml4rrieu/status/1357646193241845763?s=20>. Et le code sur Github : [https://github.com/ml4rrieu/barometre\\_science\\_ouverte\\_uvqs](https://github.com/ml4rrieu/barometre_science_ouverte_uvqs). Consultés le 5 février 2021

<sup>331</sup> « Pour une meilleure visibilité de la recherche française », Antoine Blanchard et Elifsu Sabuncu, 2015, <https://hal.archives-ouvertes.fr/hal-01251541/document/>. Consulté le 8 février 2021

<sup>332</sup> Entretien du 3 décembre 2020.

En France, l'adoption des CRIS est encore faible<sup>333</sup>, mais plusieurs solutions ont été développées, notamment avec l'appui de SCD, ces dernières années. Nous revenons sur deux d'entre elles.

À l'université de Lille, peut-être l'établissement de l'ESR le plus avancé dans ce domaine, un CRIS a été construit dans le cadre de la fusion de plusieurs établissements de champs disciplinaires différents et avec des pratiques de valorisation de la production scientifique différentes<sup>334</sup>. Plusieurs initiatives mutualisées ont participé à ce travail. Ont ainsi été mises en place une « charte de signature unique et normalisée » afin de faciliter le repérage des productions et leur valorisation dans les bases bibliographiques, puis la plateforme bibliométrique Lillometrics<sup>335</sup>. Les différents établissements détenant déjà des outils de valorisation propres, l'unification s'est faite à partir de l'existant, en créant des liens et en ajoutant les briques manquantes, plutôt qu'en édifiant un outil *ex nihilo*. Dans cette optique est née l'archive ouverte connectée à HAL Lille Open Archive (LilloA)<sup>336</sup> et l'outil Sampra, « outil de gestion et d'analyse des références », a été adopté à la suite du CHU de Lille. Ce logiciel permet « de disposer d'un recensement fiable des publications scientifiques, grâce à la récupération des références bibliographiques présentes dans PubMed et le Web of Science et à leur validation par les chercheurs »<sup>337</sup>. Un référentiel d'identité numérique permet de rassembler les informations sur les affiliations et les identifiants des chercheurs (Orcid, IdRef, IdHAL...). L'ensemble est pensé dans la logique mentionnée plus haut du « *input once, output many* », le référentiel permettant d'alimenter à la fois LilloA et les pages des chercheurs<sup>338</sup> : « Il s'agit de passer d'une logique où les données sont produites à des fins administratives à une logique d'exploitation diversifiée du même ensemble de données. » Un schéma<sup>339</sup> permet de visualiser l'architecture de ce système d'information :

---

<sup>333</sup> Le degré d'adoption des CRIS varie fortement d'un pays à l'autre : <https://dSPACECRIS.eurocris.org/>. Consulté le 28 février 2021

<sup>334</sup> « Un écosystème pour la visibilité des productions scientifiques : l'expérience de l'université de Lille », Laurence Crohem, Marie-Madeleine Géroutet et Cécile Malleret, Arabesques n° 95, 2019, <https://publications-prairial.fr/arabesques/index.php?id=1303https://publications-prairial.fr/arabesques/index.php?id=1303#ftn1>. Consulté le 9 février 2021

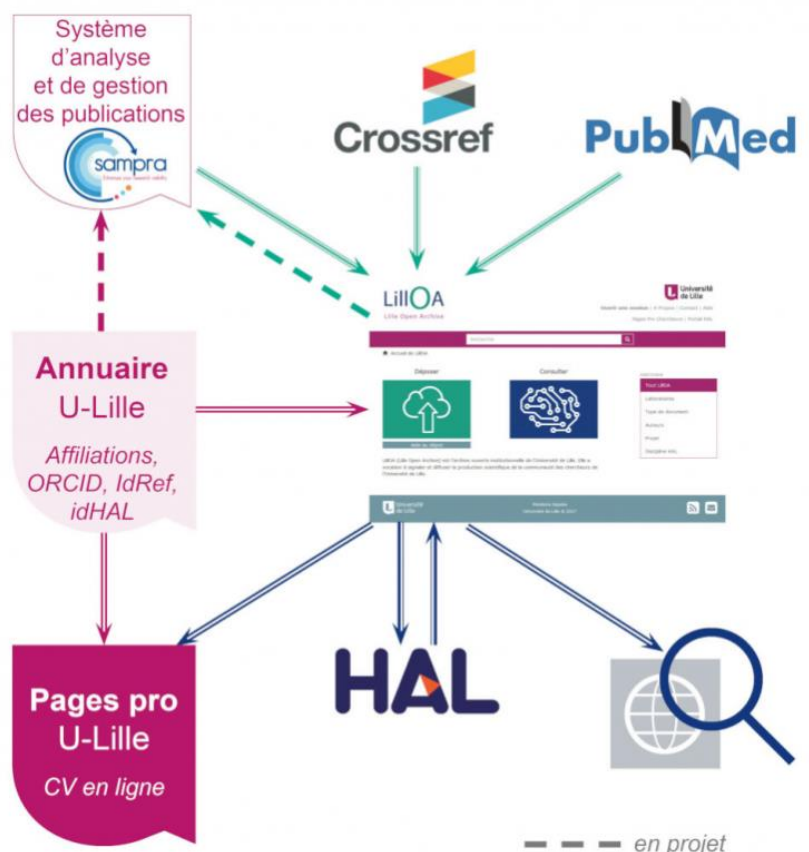
<sup>335</sup> Lillometrics, université de Lille, <https://lillometrics.univ-lille.fr/>. Consulté le 9 février 2021

<sup>336</sup> LilloA, université de Lille, <https://lilloa.univ-lille.fr/>. Consulté le 9 février 2021

<sup>337</sup> « Un écosystème pour la visibilité des productions scientifiques : l'expérience de l'université de Lille », Op. Cit.

<sup>338</sup> « Pages pro - Enseignants-chercheurs, chercheurs, enseignants », université de Lille, <https://pro.univ-lille.fr/>. Consulté le 9 février 2021

<sup>339</sup> Dans « Un écosystème pour la visibilité des productions scientifiques : l'expérience de l'université de Lille », Op. Cit.






Synthèse des flux de données dans et depuis LilloA

À Paris-Saclay, comme nous l’ont expliqué en entretien<sup>340</sup> Luc Bellier, directeur adjoint, responsable du pôle Développement et usage / Science ouverte, et Cédric Mercier, chargé de projet Données, l’université a développé un outil en interne, Bibliolabs<sup>341</sup>. À partir de 2015 dans le cadre de la Comue Paris-Saclay tout juste créée, pour faire valoir l’activité de l’université à l’international, il a d’abord été décidé de constituer un référentiel des signatures qui a permis de solidifier l’identification des quelque 300 laboratoires de l’établissement, à travers une signature normalisée que les chercheurs peuvent associer à leurs publications soumises à des éditeurs, et un alignement de différents identifiants RNSR, IdRef, Scopus, HAL. Par exemple, pour le laboratoire « Structures, propriétés et modélisation des solides » rattaché à CentraleSupélec :

<sup>340</sup> Entretien du 10 décembre 2020 avec Luc Bellier, puis du 12 février 2021 avec Luc Bellier et Cédric Mercier. Certains points de vue exprimés sont les miens, d’éventuelles erreurs sont ma responsabilité.

<sup>341</sup> Bibliolabs, université de Paris-Saclay, <https://bibliolabs.universite-paris-saclay.fr/fr>. Consulté le 9 février 2021

		Structures, propriétés et modélisation des solides	SPMS	UMR 8580	Univ. Paris-Saclay   CentraleSupélec   CNRS
RNSR:		199812908U			
ID Scopus:		60106044			
IdRef:		147968348			
Identifiant de structure HAL:		3803			
Plug in Labs ID:		914077			
Signature:		Université Paris-Saclay, CentraleSupélec, CNRS, Laboratoire SPMS, 91190, Gif-sur-Yvette, France. 			

**Notice du laboratoire « Structures, propriétés et modélisation des solides » dans le référentiel de Paris-Saclay**

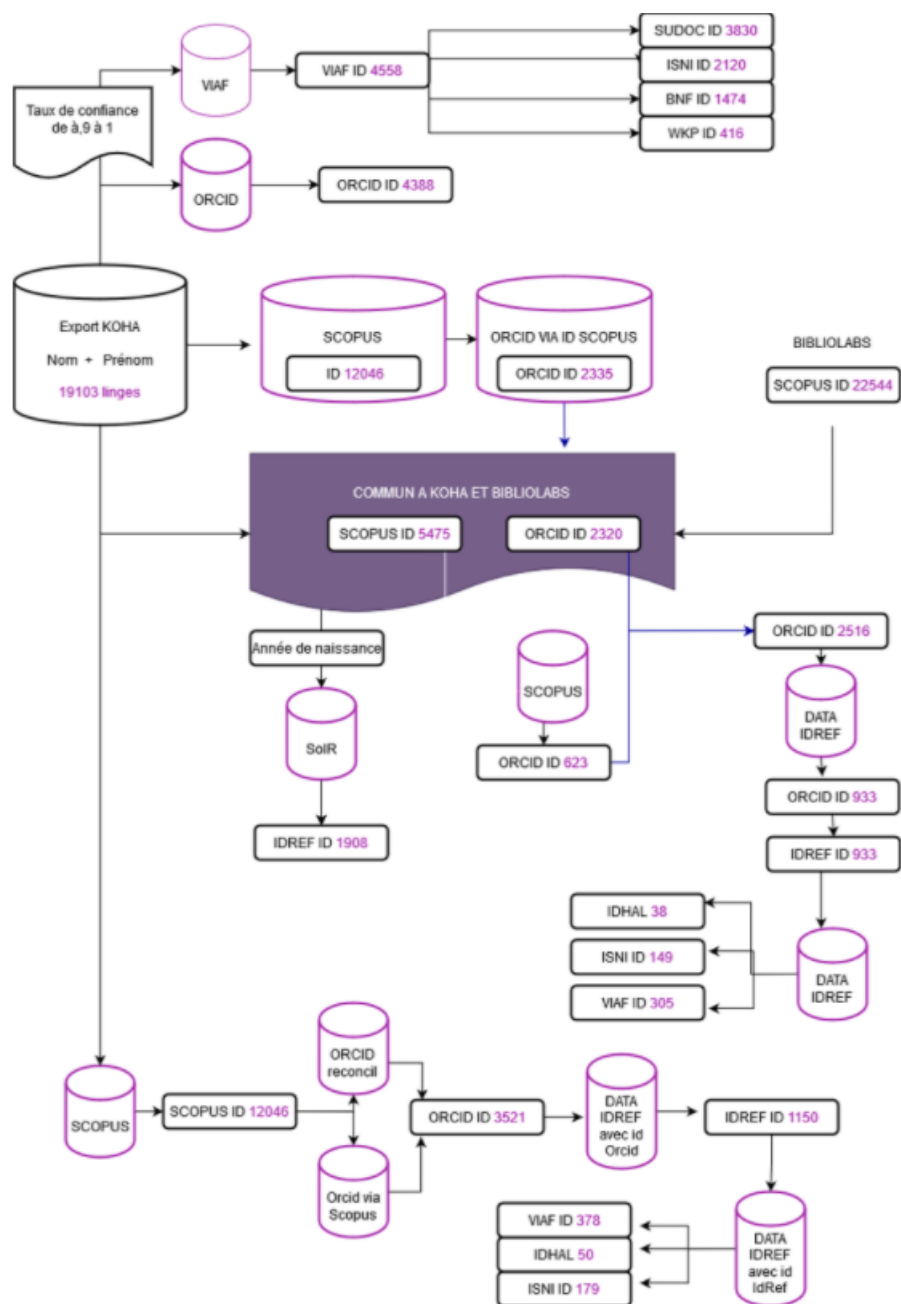
Ces laboratoires ont pu être identifiés comme parties prenantes dans la production de 13 000 à 15 000 publications recensées par an. L’outil Bibliolabs permet de récupérer les publications et leurs DOI au moyen des signatures du référentiel en interrogeant les API Scopus et WoS. Les données pivots sont celle de Scopus, qui permet une identification des laboratoires plus fine que WoS.

À partir de 2019, une brique supplémentaire a été ajoutée à la demande de la direction des relations internationales pour mesurer l’impact des financements européens sur la recherche de l’établissement. Il s’agit de recueillir dans la base des projets européens Cordis<sup>342</sup> les projets où sont impliqués des laboratoires de Saclay, ce qui permet ensuite de déterminer dans Bibliolabs les publications liées à ces projets. Puis le processus a été élargi aux jeux de données de la recherche. Cela peut avoir une utilité pour le pilotage de la recherche mais aussi pour les rapports Hcéres par exemple.

Un référentiel d’autorités des chercheurs a aussi été constitué à partir de l’annuaire, via le SIGB, avec OpenRefine. En partant des données noms, prénoms et dates de naissance de l’ensemble des personnels de l’université récupérées dans le SIGB, on a pu obtenir les ScopusID d’un certain nombre d’entre eux, avant de croiser ces ScopusID avec ceux de Bibliolabs (puisque les données bibliographiques de Bibliolabs sont récupérées de Scopus, tous les auteurs dans Bibliolabs ont au moins un ScopusID), ce qui a permis de réduire à ceux qui sont effectivement rattachés à Saclay. On a ensuite pu les aligner avec des identifiants Orcid et Viaf. Une extraction des publications de Bibliolabs a été transmise à l’Abes, qui a pu récupérer (par traitement interne) les IdRef, et à partir de ceux-ci des IdHAL et identifiants Isni et Viaf. Et, à partir de ces identifiants numériques des chercheurs, on peut récupérer leurs publications, même si la signature utilisée n’est pas celle recensée dans le référentiel. Le schéma suivant résume ce processus complexe<sup>343</sup> :

<sup>342</sup> « Cordis - Résultats de la recherche de l’UE », Commission européenne, <https://cordis.europa.eu/projects/fr>. Consulté le 9 février 2021

<sup>343</sup> Merci à Luc Bellier et Cédric Mercier de m’avoir donné accès à ce document.



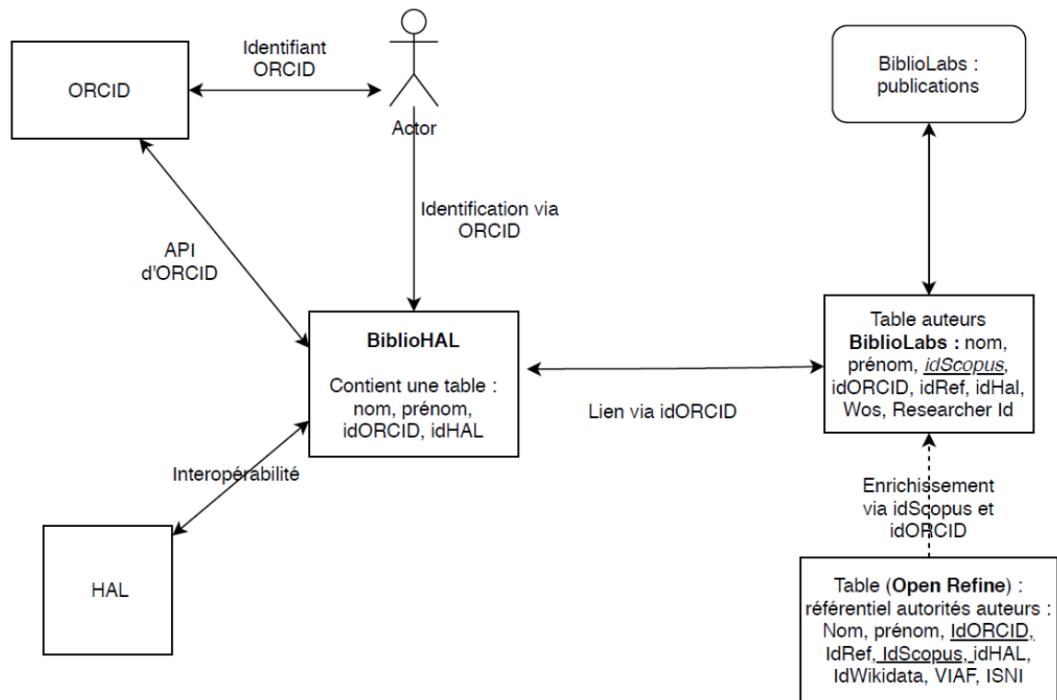
Protocole de constitution du référentiel Chercheurs de Saclay<sup>344</sup>

Un service d'aide au dépôt dans HAL sur cette base est en cours de mise en place : à partir des publications récoltées dans Bibliolabs, on peut proposer aux chercheurs de verser leurs publications dans HAL, en récupérant automatiquement les métadonnées ainsi que le texte intégral si le texte est en Open Access. Cela permet aussi de valider les liens entre publications et auteurs, en soumettant au chercheur une liste de ses publications supposées, qu'il peut valider et donner l'autorisation de verser dans HAL en fonction des co-auteurs et du statut OA de la publication. L'outil BiblioHAL, similaire à Dissemin<sup>345</sup>, mais appuyé sur un référentiel local, est en cours de développement à cet effet. Le chercheur pourra s'identifier par son Orcid à BiblioHAL, qui sera interopérable avec HAL et avec Bibliolabs, ce qui permettra de croiser les publications dans Bibliolabs, Orcid et HAL, et de déterminer lesquelles sont dans HAL et lesquelles n'y sont pas. On

<sup>344</sup> Schéma communiqué par les animateurs du projet.

<sup>345</sup> Page d'accueil, Dissemin, <https://dissem.in/>. Consulté le 9 février 2021

pourra sur cette base demander au chercheur de certifier les publications dont il est l'auteur, et lui proposer de verser dans HAL les publications qui n'y sont pas, si cela est autorisé (selon le statut OA recueilli dans Sherpa-Romeo). L'outil permettra de récupérer automatiquement les métadonnées, et le texte s'il est disponible via l'éditeur. Dans le cas contraire le chercheur devra ajouter lui-même le manuscrit accepté, mais les métadonnées seront versées automatiquement. Un schéma<sup>346</sup> permet de mieux visualiser ces flux complexes de données :



Échanges de données entre Bibliolabs, BiblioHAL, HAL et Orcid

Bibliolabs est d'une part un outil bibliométrique qui permet de valoriser la production scientifique de l'établissement, à des fins de pilotage par la gouvernance et de visibilité pour les classements, et d'autre part un outil d'analyse de la science ouverte dans le cadre du BSO local de Saclay (en récupérant les DOI dans Bibliolabs puis le statut OA dans Unpaywall à partir de ces DOI).

La direction de la recherche peut également s'appuyer sur le référentiel des laboratoires pour faire un état des lieux des collaborations à la fois au sein de Saclay et en dehors de Saclay, et à l'international, avec à chaque fois des liens vers les publications.

Cependant, un des problèmes de l'outil, en termes d'ouverture, est qu'il s'appuie essentiellement sur Scopus, et que si travailler sur ces données est autorisé, les exposer ne l'est pas. Le projet ne permet donc pas d'ouvrir les données pour le moment, même s'il sera possible de les ouvrir après un volume de traitement suffisant.

En termes de compétences et de ressources humaines, un projet de ce type requiert une bonne connaissance du paysage bibliographique, des flux du web sémantique, des identifiants, des ontologies, mais aussi une certaine maîtrise d'outils comme OpenRefine et de langages comme Python, qui nécessitent soit l'appel à des compétences extérieures,

<sup>346</sup> Merci à Luc Bellier et Cédric Mercier de m'avoir donné accès à ce document.

soit de la formation ou de l'autoformation en interne. En l'occurrence, Bibliolabs a d'abord été développé par un informaticien pendant trois ans, qui a pu former l'administrateur actuel, de profil non-informaticien. Les traitements dans OpenRefine ont été réalisés par une bibliothécaire assistante spécialisée, dotée d'un diplôme en informatique. Il serait souhaitable, pour permettre aux professionnels de l'IST de s'impliquer plus efficacement dans des projets de ce type, de développer la formation à OpenRefine et à Python pour améliorer les traitements et les réaliser de façon plus industrielle. De façon générale, ces formations plus techniques pourraient être pertinentes dans les formations initiales et continues des bibliothécaires, ce qui favoriserait le développement de ce type de traitements des données.

## Bilan

Dans ce chapitre, nous avons examiné une série de projets permettant la curation selon des procédures semi-automatiques des données des éditeurs et de l'ESR. La gestion des métadonnées des éditeurs est plutôt du ressort des agences bibliographiques, et permet de pallier dans une certaine mesure les problèmes de qualité de ces métadonnées. Le travail sur les données de l'ESR concerne les agences, certains services ministériels mais aussi les établissements eux-mêmes, dans une démarche d'analyse, de pilotage et d'exposition de leur production scientifique. Ce travail implique la mise en place de politiques d'identifiants, de référentiels, d'alignements, qui requièrent des compétences spécifiques (et donc de la formation) et des moyens. Ces outils doivent permettre, au moins s'ils sont largement partagés au sein de la communauté de l'ESR, une amélioration des référentiels à l'échelle nationale, voire au-delà, avec une meilleure présence de la production française dans les bases de données, les référentiels et registres internationaux (Crossref, Orcid, Wikidata...).



## CONCLUSION

---

Une multitude d'acteurs interviennent de manière cruciale pour la gestion globale des métadonnées de publications scientifiques : les agences de gestion des identifiants et les grandes bases de données au niveau mondial, les agences bibliographiques au niveau national, dont le travail s'appuie sur le travail des professionnels de l'IST dans leur réseau, au niveau local. Tous les professionnels de l'IST n'ont pas vocation à s'investir dans la création d'outils d'intelligence artificielle, mais chaque SCD peut contribuer à son échelle à rendre plus solides les liens entre les entités décrites (en mettant en place une politique rigoureuse d'identifiants chercheurs et structures dans leur université, par exemple) ou à améliorer les résultats des algorithmes de traitement, qui ne peuvent jamais être intégralement automatisés (dédoublonnage, alignements incertains...). Ces orientations renouvellent le signalement des ressources et rappellent l'idée de contrôle bibliographique universel (CBU) né dans les années 1970 à propos du catalogage. Mais cette idée de CBU prend un nouveau souffle dans le cadre du web sémantique, avec un partage sans équivalent des données, qui ne rend pas moins nécessaire le travail de curation sur ces données, mais favorise la mutualisation de ce travail. Comme l'affirmaient dès 2014 Gildas Illien et Françoise Bourdon dans un article sur la nouvelle ère du CBU dans le cadre du web sémantique, « la constitution de registres collaboratifs d'autorités associés à des identifiants standardisés constitue de [leur] point de vue un pilier fondamental du nouveau contrôle bibliographique universel »<sup>347</sup>. Chaque établissement de l'ESR peut être pensé comme déléataire d'une mission de signalement des productions et autorités locales, et les agences bibliographiques d'une mission de signalement pour les productions des éditeurs nationaux, et de fournitures de services aux établissements pour qu'ils puissent effectivement signaler la production locale dans de bonnes conditions : référentiels, infrastructures techniques, formations, négociations avec les éditeurs pour la mise à disposition de leurs métadonnées, traitement de ces métadonnées, centralisation des données locales dans un « pot commun » national, lui-même mis à disposition d'acteurs au niveau international, etc.

On peut tenter, une fois qu'on a admis le constat d'un rôle actuel et potentiel important des professionnels de l'IST dans cette dynamique, se demander quelles évolutions pourraient être mises en place. On peut distinguer trois grandes évolutions souhaitables : 1) une montée en compétence et une acculturation plus poussée des bibliothécaires aux données, aux métadonnées et à leur traitement (automatique ou non), à partir de savoir-faire déjà existants dans les établissements ; 2) une rationalisation de l'organisation des institutions nationales pilotant ces travaux, qui aujourd'hui se superposent voire se font concurrence ; 3) un plus grand investissement de la France dans les infrastructures internationales de la science ouverte.

1) Nous avons orienté notre propos dans des directions diverses (identifiants, pépinières de revues, archives ouvertes, systèmes d'informations recherche, traitements automatiques des données, etc.), qui en réalité s'articulent entre elles et se complètent. Car ces différents chantiers reposent sur des savoirs et savoir-faire similaires et déjà répandus dans les SCD car ils correspondent aux compétences traditionnelles des bibliothécaires : la gestion des autorités, l'indexation et le signalement selon des standards, le recours à des protocoles d'échange d'informations, tout cela fait partie de la panoplie des bibliothécaires depuis longtemps, et s'applique en s'adaptant aux nouveaux

---

<sup>347</sup> « À la recherche du temps perdu, retour vers le futur : CBU 2.0 », Gildas Illien et Françoise Bourdon, 2014, <http://library.ifla.org/956/1/086-illien-fr.pdf>. Consulté le 20 février 2021

enjeux, comme le notait Aline Le Provost dans le n° 97 d'*Arabesques* : « La gestion des jeux de données numériques est devenue un enjeu incontournable de la science ouverte. La curation de ces données intègre des tâches de sélection, de vérification, de normalisation, de structuration ou encore d'enrichissement, indispensables pour publier des données ré-exploitable. Dans ce domaine, les bibliothécaires ont pour atout leur expérience collective et partagée de tout ou partie de ces tâches. Avec l'ouverture des catalogues et la diffusion de leurs métadonnées, ils sont déjà devenus des gestionnaires de données potentiellement utilisables par la communauté des chercheurs »<sup>348</sup>. Pour autant, ces compétences restent l'apanage d'une partie des agents, même si elles ont tendance à s'étendre.

Les référentiels métiers de l'IST sont à cet égard encore discrets sur les compétences précises identifiées ici<sup>349</sup>. Référens III fait apparaître les « méthodes de gestion des données », sans plus de précisions, pour la fiche « Responsable des ressources et de l'ingénierie documentaire »<sup>350</sup>. La fiche « Chargé-e du traitement des données scientifiques »<sup>351</sup> est plus précise, faisant apparaître les éléments « Diversification des supports d'information (accroissement du volume de données, augmentation de leur complexité et accès) », « Évolution des techniques du web de données (libre accès, web sémantique) et de leur accompagnement », « Développement de la dématérialisation des ressources documentaires, de l'information en ligne et des procédures » et « Développement des réseaux nationaux et internationaux et normalisation des pratiques », mais seulement dans les « facteurs d'évolution à moyen terme ». Elles font pourtant partie intégrante des compétences nécessaires à nombre de structures documentaires, ce qui se traduit d'ores et déjà dans de nombreuses fiches de poste.

Dans l'optique d'une évolution des compétences existantes, notamment liées au catalogage, pour répondre aux nouveaux enjeux (archives institutionnelles, services aux chercheurs, ressources électroniques...), l'Abes a impulsé en 2018 l'évolution, décrite dans *Arabesques* par Laurent Piquemal, du rôle de coordinateur Sudoc vers celui de « référent métadonnées »<sup>352</sup>, avec une montée en compétences et une vision plus stratégique et élargie à l'ensemble de l'établissement, consistant à « aller à la rencontre des acteurs collaborant, comme lui, au système d'information ; partager sa connaissance des workflows de données et sa maîtrise des formats ; transmettre les compétences acquises en termes d'organisation de chantiers qualités ; valoriser son expérience en termes d'administration et de correction de données ». Cette évolution est donc symptomatique d'un élargissement du domaine des données et des métadonnées, qui ne sont plus restreintes à leur pré carré du catalogage, mais innervent l'ensemble de l'activité de la structure documentaire<sup>353</sup>. Il faut donc s'assurer que l'ensemble des agents

<sup>348</sup> « La curation, un enjeu pour la gestion des données numériques », Aline Le Provost, *Arabesques* n° 97, 2020, <https://publications-prairial.fr/arabesques/index.php?id=1793>. Consulté le 9 février 2021

<sup>349</sup> Voir « Référentiels métiers, référentiels de compétences : bilan et perspectives », Nathalie Marcerou-Ramel, *Bulletin des bibliothèques de France* (BBF), 2017, n° 13, p. 8-18, <https://bbf.enssib.fr/consulter/bbf-2017-13-0008-001> ISSN 1292-8399. Consulté le 5 mars 2021

<sup>350</sup> « F1A41 - Responsable des ressources et de l'ingénierie documentaire », Référens III, Mesri, [https://data.enseignementsuprecherche.gouv.fr/pages/fiche\\_emploi\\_type\\_referens\\_iii\\_itrf/?refine.referens\\_id=F1A41#top](https://data.enseignementsuprecherche.gouv.fr/pages/fiche_emploi_type_referens_iii_itrf/?refine.referens_id=F1A41#top) Consulté le 5 mars 2021

<sup>351</sup> « F2A43 - Chargé-e du traitement des données scientifiques », Référens III, Mesri, [https://data.enseignementsup-recherche.gouv.fr/pages/fiche\\_emploi\\_type\\_referens\\_iii\\_itrf/?refine.referens\\_id=F2A43#top](https://data.enseignementsup-recherche.gouv.fr/pages/fiche_emploi_type_referens_iii_itrf/?refine.referens_id=F2A43#top). Consulté le 5 mars 2021

<sup>352</sup> « L'évolution de la fonction de Coordinateur Sudoc : une opportunité pour la construction d'une politique globale des métadonnées », Laurent Piquemal, *Arabesques* n° 97, 2020, <https://publications-prairial.fr/arabesques/index.php?id=1792>. Consulté le 9 février 2021

<sup>353</sup> Une même analyse de la continuité entre catalogage Marc et traitement de métadonnées non-Marc, dans un contexte américain, est fait par un récent article : « The Roles of Cataloging vs. Non-Cataloging Librarians and Staff in Non-MARC Metadata-Related Workflows: A Survey of Academic Libraries in the United States », Jeannette Ho, *Cataloging & Classification Quarterly*, 2021, p. 1-33, <https://doi.org/10.1080/01639374.2020.1863889>. Consulté le 15 février 2021

développent une culture de la donnée adaptée à leur fonction, tout en s'assurant que certains agents développent une véritable expertise à la fois technique et institutionnelle, une compréhension profonde des enjeux des métadonnées à l'heure de la science ouverte, pour pouvoir mettre en place des chantiers ambitieux.

Ainsi, s'il n'est pas possible ni souhaitable que tout bibliothécaire maîtrise Python ou soit capable d'écrire lui-même des scripts, il est important que de telles compétences, encore rares, se développent, afin de pouvoir mener par exemple des adaptations locales de projets menés dans les institutions, comme on a pu le voir par exemple dans le cas des baromètres de la science ouverte locaux. Par ailleurs, l'inscription dans les formations initiales et continues des bibliothécaires d'une introduction aux techniques de curation, de dédoublement, sous la forme par exemple d'une initiation à des outils comme OpenRefine, pourraient être bienvenue<sup>354</sup>. De même, le renforcement de formations autour du web sémantique pourraient être utiles afin que les professionnels de l'IST partagent une connaissance des enjeux et des outils autour des *linked open data* (XML, RDF, identifiants, ontologies, Sparql, mais aussi connaissance des grands acteurs de la science ouverte, du cadre juridique, etc.). Notons que de telles formations générales permettraient en outre une appréhension plus facile des enjeux et mécanismes de la Transition bibliographique, du modèle Ifla-LRM, de RDA, etc.

2) On a vu que les acteurs de l'écosystème des métadonnées étaient multiples et que leurs actions pouvaient parfois se superposer quand elles n'étaient pas tout simplement en concurrence. Une rationalisation des initiatives à l'échelle de l'ESR permettrait d'améliorer l'efficacité, en mutualisant davantage les moyens. Nous avons signalé la création récente d'un poste administratrice des données de l'ESR, occupé par Mme Isabelle Blanc<sup>355</sup>, dont l'action doit permettre d'aller en ce sens. Les institutions qui sont en charge de parties plus ou moins importantes du travail sur les données de l'ESR sont multiples et leur intégration pourrait être poussée plus loin, comme le suggère Benoît Lecoq, inspecteur général de l'Éducation, du Sport et de la Recherche, dans un récent numéro d'*Arabesques* : « *Il nous semble qu'aujourd'hui le périmètre des compétences de l'Abes demanderait à être étendu et clarifié en sorte de fédérer aussi bien les établissements publics dédiés aux services documentaires (parmi lesquels l'Inist) que la gerbe de services et d'initiatives portées par des organisations aux statuts les plus divers – et parfois fragiles – : le consortium Couperin, le Centre pour la communication scientifique directe (CCSD), le GIS Collex-Persée, le Comité pour la science ouverte (CoSO), etc.* »<sup>356</sup> En effet, la mise en place d'une politique unifiée autour d'identifiants et de référentiels uniques (pour les chercheurs, les institutions...) pourrait être favorisée par une rationalisation de la gouvernance, qui génère une multiplication des équipes et des budgets, ce qui peut créer des intérêts parfois divergents à court terme, bien que les objectifs à long terme soient les mêmes. Pour les autorités bibliographiques, le lancement prochain du Fichier national d'entités en coproduction entre les réseaux du Sudoc et de la BnF, les bibliothèques publiques, voire les centres de documentation, les archives, les institutions culturelles en France et dans l'espace francophone, est aussi un pas vers une mutualisation renforcée.

3) Enfin, les experts français des métadonnées bibliographiques sont encore peu

<sup>354</sup> Même si les performances d'OpenRefine sont parfois critiquées, notamment pour les gros volumes de données. Voir par exemple le tweet de Gautier Poupeau, de l'INA, du 9 février 2021 : <https://twitter.com/lespetitescases/status/1359188211625562118?s=20>. Consulté le 12 février 2021

<sup>355</sup> Profil LinkedIn d'Isabelle Blanc, Chief Data Officer chez Ministère en charge de l'enseignement supérieur et de la recherche (France), <https://www.linkedin.com/in/isabelleblancatala/?originalSubdomain=fr>. Consulté le 9 février 2021

<sup>356</sup> Benoît Lecoq, « Les Trente Glorieuses de la connaissance ? », *Arabesques* n° 100, 2021, <https://publications-prairial.fr/arabesques/index.php?id=2290>. Consulté le 9 février 2021

investis dans les instances internationales de gouvernance des identifiants pérennes et de la science ouverte. Nous avons signalé que le Comité pour la science ouverte a lancé le 1<sup>er</sup> février 2021 un appel à manifestation d'intérêt pour la construction d'un Réseau d'experts internationaux de la Science ouverte (Reiso). Cela pourrait permettre à moyen terme d'asseoir une meilleure représentation de la France et des points de vue qu'elle porte dans les institutions internationales. Dans l'immédiat, elle témoigne au moins d'une prise de conscience et d'une volonté de s'inscrire dans une dynamique plus large. Cette dynamique était déjà visible dans les orientations du Plan national pour la science ouverte, qui a abouti à la création d'un consortium Orcid France fin 2019, témoignant d'une volonté d'investissement dans la dynamique des identifiants pérennes et d'une politique coordonnée à l'échelle nationale.

Montée en compétences des agents de l'IST française, rationalisation de la gouvernance à l'échelle nationale et implication renforcée dans le web sémantique mondial nous paraissent donc les enjeux structurants des prochaines années pour un passage réussi à l'ère des données ouvertes liées.

## ANNEXE

### LES OUTILS TECHNIQUES DES MÉTADONNÉES DES RESSOURCES SCIENTIFIQUES

Il nous paraît important de présenter, au moins de manière succincte, l'infrastructure technique qu'il sera bienvenu qu'un professionnel de l'IST connaisse pour participer à la gestion des métadonnées scientifiques. Le panorama qui suit n'est pas exhaustif. Il s'agit d'une version étendue de la section « Un arsenal d'outils pour décrire les ressources scientifiques de façon standard » dans notre chapitre 1.

#### Quelle syntaxe pour encoder les métadonnées ?

Pour gérer efficacement les métadonnées d'une ressource, et les rendre partageables, il est indispensable de disposer de formats standards, c'est-à-dire de formats et de protocoles partagés permettant l'interopérabilité, lisibles par des machines. Ces formats existent à différents niveaux, à commencer par la structure de leur encodage.

Un standard omniprésent est le langage de balisage XML (eXtensible Markup Language), permettant d'organiser hiérarchiquement des éléments à partir d'une racine. XML est un « format ouvert et standardisé pour exporter et importer des données structurées entre deux systèmes d'informations »<sup>357</sup>. Dans un cadre bibliographique, il permet de rattacher un ensemble d'informations à une œuvre, par exemple son titre et son auteur, puis de lier à cet auteur les informations qui le concernent, par itération de la structure hiérarchique. Par exemple, la séquence XML suivante établit que l'œuvre *La Tempête* est de type « Pièce de théâtre » et qu'elle a été écrite par un dramaturge. Puis, que le nom de ce dramaturge est William Shakespeare et qu'il est né à Stratford upon Avon<sup>358</sup> :

```
<?xml version="1.0" encoding="UTF-8"?>
<work type="play">
  <workName>The Tempest</workName>
  <writtenBy>
    <playwright>
      <playwrightName>William Shakespeare</playwrightName>
      <bornInPlace>Stratford Upon Avon</bornInPlace>
    </playwright>
  </writtenBy>
</work>
```

Ce code ne fait appel à aucun élément extérieur, en dehors de la version de XML « 1.0 » et du format d'encodage UTF-8<sup>359</sup>. Mais pour garantir davantage de convergences

<sup>357</sup> *Introduction aux humanités numériques*, Max De Wilde, Florence Gillet, Simon Hengchen, Seth van Hooland, 2016, p. 63.

<sup>358</sup> « Understanding Metadata : What is metadata, and what is it for? », 2017, Jenn Riley / Niso, p. 13, [https://groups.niso.org/apps/group\\_public/download.php/17446/Understanding%20Metadata.pdf](https://groups.niso.org/apps/group_public/download.php/17446/Understanding%20Metadata.pdf). Consulté le 1<sup>er</sup> février 2021

<sup>359</sup> « UTF-8 », Wikipédia, <https://fr.wikipedia.org/wiki/UTF-8>. Consulté le 1<sup>er</sup> février 2021

entre bases de données, un document XML peut avoir recours à des espaces de noms (ou *namespaces*), c'est-à-dire des espaces virtuels externes de termes et leurs définitions dans lesquels le document XML puise des propriétés au sens univoque, partagé par tous les autres acteurs qui se référeront au même espace de noms. Par exemple, le vocabulaire Dublin Core (que nous détaillons ci-dessous) permet de déterminer de façon standard de quoi il est question si l'on mentionne la métadonnée « Creator » dans le document XML (en l'occurrence : « Nom de la personne, de l'organisation ou du service responsable de la création du contenu de la ressource »).

**JSON** (JavaScript Object Notation) est un langage permettant de « *représenter de l'information structurée* », comme XML<sup>360</sup>. Mais « *par rapport à XML il n'y a pas de notion d'espace de noms. C'est un modèle plus souple* »<sup>361</sup>. Voici un exemple de fichier JSON :

```

1      {
2          "firstName": "John",
3          "lastName" : "Smith",
4          "age"      : 25,
5          "address"  :
6          {
7              "streetAddress": "21 2nd Street",
8              "city"       : "New York",
9              "state"      : "NY",
10             "postalCode" : "10021"
11         },
12         "phoneNumber":
13         [
14             {
15                 "type" : "home",
16                 "number": "212 555-1234"
17             },
18             {
19                 "type" : "fax",
20                 "number": "646 555-4567"
21             }
22         ]
23     }

```

## Standards de métadonnées

L'interopérabilité implique également des *schémas de métadonnées* partagés, spécifiant les informations obligatoires ou optionnelles que doivent renseigner les acteurs

<sup>360</sup> « JavaScript Object Notation », Wikipédia, [https://fr.wikipedia.org/wiki/JavaScript\\_Object\\_Notation](https://fr.wikipedia.org/wiki/JavaScript_Object_Notation). Consulté le 9 février 2021

<sup>361</sup> « Structuration des données: XML et Json », CentraleSupélec – Département d'informatique, <https://wdi.supelec.fr/appliouaibe/Cours/XML>. Consulté le 9 février 2021

en charge des métadonnées d'une ressource (chercheurs, curateurs, éditeurs...). Ces schémas répondent à des contraintes différentes suivant les usages qui en sont faits.

D'abord, l'**ISBD** (International Standard Bibliographic Description), élaboré par l'Ifla, est un « ensemble normatif de règles validées au niveau international, pour la description bibliographique de toute ressource publiée existant dans les bibliothèques, quel qu'en soit le support »<sup>362</sup>, qui :

- « précise les **éléments requis** pour une description bibliographique établie par une agence bibliographique nationale ;
- prescrit **leur ordre** de présentation, en les répartissant entre 8 zones cohérentes ;
- définit **une ponctuation** pour les délimiter, cette ponctuation étant utilisée comme **un codage** ;
- donne des règles pour **leur transcription** à partir de **sources d'information** clairement identifiées »<sup>363</sup>.

Fondé sur l'ISBD, le format Machine-Readable Cataloging (**Marc**) permet l'échange de métadonnées entre les catalogues. Lancé en 1968 par la Library of Congress, il renvoie en réalité à une famille de formats, qui diffèrent suivant les pays notamment : Marc21 pour les pays anglo-saxons, et Unimarc pour le reste du monde, par exemple la France (mais c'est encore un autre format, Intermarc, qui est utilisé à la BnF).

Marc se présente comme une succession de champs (auteur, titre, etc.) dont le type d'information est spécifié afin qu'une machine puisse la récupérer correctement, et labellisés par des étiquettes (*tags*). Voici quelques étiquettes données en exemple dans la documentation de la Library of Congress, donc pour Marc 21<sup>364</sup> :

- « 010 tag marks the Library of Congress Control Number (LCCN)*
- 020 tag marks the International Standard Book Number (ISBN)*
- 100 tag marks a personal name main entry (author)*
- 245 tag marks the title information (which includes the title, other title information, and the statement of responsibility)*
- 250 tag marks the edition*
- 260 tag marks the publication information*
- 300 tag marks the physical description (often referred to as the "collation" when describing books)*
- 490 tag marks the series statement*
- 520 tag marks the annotation or summary note*
- 650 tag marks a topical subject heading*
- 700 tag marks a personal name added entry (joint author, editor, or illustrator) »*

<sup>362</sup> « ISBD (International Standard Bibliographic Description) », BnF, <https://www.bnf.fr/fr/isbd-international-standard-bibliographic-description>. Consulté le 9 février 2021

<sup>363</sup> Ibid.

<sup>364</sup> « What is a Marc record, and why is it important? », Library of Congress, <https://www.loc.gov/marc/umb/um01to06.html>. Consulté le 9 février 2021

Plus généralement, les champs sont organisés de la façon suivante<sup>365</sup> :

- « 0XX *Control information, numbers, codes*
- 1XX *Main entry*
- 2XX *Titles, edition, imprint (in general, the title, statement of responsibility, edition, and publication information)*
- 3XX *Physical description, etc.*
- 4XX *Series statements (as shown in the book)*
- 5XX *Notes*
- 6XX *Subject added entries*
- 7XX *Added entries other than subject or series*
- 8XX *Series added entries (other authoritative forms)* »

Le besoin d'un format permettant l'échange d'informations au-delà des bibliothèques a conduit à la création en 1995 du schéma de métadonnées **Dublin Core**, déterminant d'abord le Dublin Core Element Set, constitué de 15 éléments (titre, créateur, sujet, description, éditeur, contributeur, date, type, format, identifiant, source, langue, relation, couverture, gestion des droits) permettant de décrire des ressources très diverses et de relier ces descriptions entre elles. Un ensemble supplémentaire DC Terms<sup>366</sup> a ensuite vu le jour pour élargir la gamme de propriétés prises en charge par ce format. Mais ce format peut faire perdre des informations en raison de sa grande simplicité<sup>367</sup>.

La **TEI (Text Encoding Initiative)**<sup>368</sup> permet non seulement d'encoder les métadonnées d'une ressource, via un module spécifique, <teiHeader>, mais aussi de structurer le texte lui-même, d'y introduire de la sémantique de façon à ce que le texte contienne des métadonnées en interne, par exemple pour signifier qu'un mot est un nom de lieu, qu'il désigne Paris en France et non Paris au Texas, ou encore que tel texte est un poème, et ses parties des quatrains ou des sonnets. De nombreux éditeurs structurent les textes des articles selon ce langage de balisage, ce qui permet de disposer d'une ressource sémantiquement riche<sup>369</sup>.

**EAD** est un standard, inspiré de la TEI, utilisé pour la description d'archives, de manuscrits, de cartes, ou d'« instruments de recherche » (c'est-à-dire « un ouvrage, un fichier ou une base de données qui décrit le contenu d'un ensemble de documents conservés par un service d'archives »<sup>370</sup>) basé sur XML. EAD implique les métadonnées suivantes :

<sup>365</sup> Ibid.

<sup>366</sup> « DCMI Metadata Terms », DCMI, <https://dublincore.org/specifications/dublin-core/dcmi-terms/#http%3a%2f%2fpurl.org%2fdc%2felements%2f1.1%2fdescription>. Consulté le 9 février 2021

<sup>367</sup> Voir dans cette vidéo des journées de l'Abes 2018, pour un exemple sur le moissonnage de Calames, où Dublin Core fait perdre certaines informations propres à l'EAD : <http://www.canalc2.tv/video/14965>, vers 1'. Consulté le 9 février 2021

<sup>368</sup> « Text Encoding Initiative », Wikipedia, [https://en.wikipedia.org/wiki/Text\\_Encoding\\_Initiative](https://en.wikipedia.org/wiki/Text_Encoding_Initiative). Consulté le 1<sup>er</sup> février 2021

<sup>369</sup> Par exemple, la revue *Arabesques* de l'Abes structurent ainsi ses textes mis en ligne : à l'adresse <https://publications-prairial.fr/arabesques/index.php?id=2359&file=1>, on trouve ainsi la version XML-TEI de « Songez que du haut de ce numéro 100 d'*Arabesques*, 25 ans d'histoire des bibliothèques vous contemplent », l'édition de David Aymonin du numéro 100 de l'Abes : <https://publications-prairial.fr/arabesques/index.php?id=2288>. Consultés le 1<sup>er</sup> février 2021

<sup>370</sup> « Instrument de recherche », Wikipédia, [https://fr.wikipedia.org/wiki/Instrument\\_de\\_recherche](https://fr.wikipedia.org/wiki/Instrument_de_recherche). Consulté le 9 février 2021



- « des éléments d'identification et d'informations relatifs à l'instrument de recherche lui-même : contexte de création de l'instrument de recherche, mentions de responsabilité intellectuelle, informations administratives, techniques et de gestion ;
- des éléments d'identification et de description du fonds ou de la collection : identifiant, localisation, producteur et contexte de production, caractéristiques matérielles, contenu et organisation, modalités d'accès ;
- des éléments de description de chacun des composants et sous-composants ;
- des éléments d'informations complémentaires : documents en relation, références bibliographiques ;
- des éléments d'indexation. »<sup>371</sup>

**Bibframe** (Bibliographic Framework), est un modèle de données qui pourrait (ou pas) remplacer à terme Marc 21. Son but est d'adapter la description bibliographique aux données liées<sup>372</sup>. Il organise la description selon trois niveaux : Œuvres, Instance et Item, ainsi définis par la Library of Congress :

« *Work. The highest level of abstraction, a Work, in the Bibframe context, reflects the conceptual essence of the cataloged resource: authors, languages, and what it is about (subjects).*

• *Instance. A Work may have one or more individual, material embodiments, for example, a particular published form. These are Instances of the Work. An Instance reflects information such as its publisher, place and date of publication, and format.*

• *Item. An item is an actual copy (physical or electronic) of an Instance. It reflects information such as its location (physical or virtual), shelf mark, and barcode.* »<sup>373</sup>

À cela s'ajoutent trois concepts clés, reliés aux classes fondamentales : Agent, Sujet et Événement. Bibframe organise donc l'information bibliographique autour de différents niveaux d'abstraction, là où Marc agrège des informations sur l'Œuvre, entité abstraite, et sa réalisation dans un exemplaire particulier<sup>374</sup>, même si Marc distingue les notices bibliographiques des notices d'exemplaires localisant la ressource. Bibframe fait en outre un usage plus systématique des identifiants (de personne, de lieu, etc.) que ne le fait Marc.

**RDA** (Ressource : Description et Accès) est un code de catalogage basé sur le modèle Ifla-LRM, dont l'objectif est, dans le cadre de la « transition bibliographique »<sup>375</sup>, « d'inscrire les catalogues de bibliothèques dans l'univers du web et, dans la pratique du catalogage, de prendre en compte la réalité de l'information numérique et l'apport de l'échange de métadonnées pour la création comme pour la diffusion de l'information bibliographique »<sup>376</sup>.

Le modèle Ifla-LRM fait suite aux modèles FRBR, FRAD et FRSAD et « permet d'améliorer la présentation des catalogues à travers des entités telles que Œuvre,

<sup>371</sup> EAD : Encoded Archival Description, BnF, <https://www.bnf.fr/fr/ead-encoded-archival-description>. Consulté le 1<sup>er</sup> février 2021

<sup>372</sup> « Overview of the Bibframe 2.0 Model », Library of Congress, <https://www.loc.gov/bibframe/docs/bibframe2-model.html>. Consulté le 9 février 2021

<sup>373</sup> Ibid.

<sup>374</sup> « Bibframe Frequently Asked Questions », Library of Congress, <https://www.loc.gov/bibframe/faqs/>. Consulté le 9 février 2021

<sup>375</sup> Voir Arabesques n° 87, 2017, <https://publications-prairial.fr/arabesques/index.php?id=209>. Voir aussi, sur la « FRBRisation du Sudoc », la vidéo : <https://vimeo.com/415094781>. Et aussi la vidéo de l'Abes sur la Transition bibliographique : <https://vimeo.com/419864690>. Consultés le 5 mars 2021

<sup>376</sup> « Code de catalogage RDA », BnF, <https://www.bnf.fr/fr/code-de-catalogage-rda>. Consulté le 5 mars 2021

Expression, Manifestation, Item, Agent. Ces entités plus visibles sur le web ouvrent ainsi la voie vers une interopérabilité accrue des données des catalogues de bibliothèques »<sup>377</sup>. L'idée générale de cette démarche est que « des catalogues FRBRisés permettront de s'extraire de la logique de document pour fournir de l'information, à savoir des données d'autorité sur des personnes, des organisations, des concepts, des événements, etc. »<sup>378</sup>.

**Jats** (Journal Article Tag Suite)<sup>379</sup> est un format utilisé pour décrire la littérature scientifique, développé par Niso. Il est utilisé par de nombreux éditeurs internationaux à la place d'un format propriétaire spécifique à l'éditeur<sup>380</sup>. Nombre de livraisons Istex sont faites sous ce format par exemple, et donc sous la forme de « documents XML standardisés, riches et précis, contenant à la fois le texte et les métadonnées »<sup>381</sup>. Ce format connaît un certain succès, dû à son utilisation dans PubMed Central. En effet, l'obligation de publier en Open Access, dans PubMed Central, le texte intégral des publications financées par les National Institutes of Health (NIH) américains, a nécessité la généralisation d'un standard.

Sous format Jats a aussi été développée par Casrai la taxonomie CrediT<sup>382</sup> (Contributor Roles Taxonomy), détaillant 14 rôles qui peuvent être joués par des acteurs de la recherche scientifique, au-delà du trop général « Auteur » ou « Créateur » / « Creator » en Dublin Core par exemple (Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources Software Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing). Ce dispositif permet de créditer davantage d'acteurs, et pas seulement le ou les auteurs d'une ressource scientifique, au sens strict du terme.

Le schéma **Mods** (Metadata Object Description Schema) est « un vocabulaire XML de description bibliographique développé en 2002 par le Bureau des normes et du développement du réseau de la Bibliothèque du Congrès »<sup>383</sup>, à la complexité intermédiaire entre Marc et Dublin Core. Mods « est particulièrement intéressant dans le cadre de projets de description de documents numérisés car il contient des éléments permettant de renseigner les données relatives à la description d'une version numérisée d'un document »<sup>384</sup>. De plus, un format parallèle à Mods, utilisé pour les données bibliographiques, Metadata Authority Description Schema (Mads), permet de modéliser les données d'autorité<sup>385</sup>.

On peut encore citer **Mets**<sup>386</sup> (Metadata Encoding and Transmission Standard), servant « à exprimer des métadonnées de nature diverse portant sur un document numérique, dans le but de faciliter son échange, sa gestion et sa préservation »<sup>387</sup>, ou

<sup>377</sup> Ibid.

<sup>378</sup> « La Transition bibliographique en France », Transition bibliographique, <https://www.transition-bibliographique.fr/enjeux/position-francaise-rda/>

<sup>379</sup> « Journal Article Tag Suite », Wikipédia, [https://en.wikipedia.org/wiki/Journal\\_Article\\_Tag\\_Suite](https://en.wikipedia.org/wiki/Journal_Article_Tag_Suite). Consulté le 9 février 2021

<sup>380</sup> « Atomes crochus. Les métadonnées des éditeurs et l'Abes », Yann Nicolas, in *Vers de nouveaux catalogues*, sous la direction d'Emmanuelle Bermès (2016)

<sup>381</sup> Ibid.

<sup>382</sup> « CRediT – Contributor Roles Taxonomy », Casrai, <https://casrai.org/credit/>. Consulté le 9 février 2021

<sup>383</sup> « Metadata Object Description Schema », Wikipédia, [https://fr.wikipedia.org/wiki/Metadata\\_Object\\_Description\\_Schema](https://fr.wikipedia.org/wiki/Metadata_Object_Description_Schema) Consulté le 9 février 2021

<sup>384</sup> « Mods : présentation du format », BnF, <https://www.bnf.fr/fr/mods-presentation-du-format>. Consulté le 9 février 2021

<sup>385</sup> Ibid.

<sup>386</sup> « Mets », Library of Congress, <http://www.loc.gov/standards/mets/mets-home.html>. Et aussi : « Metadata encoding and transmission standard: primer and reference manual », Library of Congress, <http://www.loc.gov/standards/mets/METSPrimer.pdf#page=86&zoom=100,92,601>. Consultés le 1<sup>er</sup> février

<sup>387</sup> « Mets : Metadata Encoding and Transmission Standard », BnF, <https://www.bnf.fr/fr/mets-metadata-encoding-and-transmission-standard>. Consulté le 9 février 2021

**Premis**<sup>388</sup> (Preservation Metadata: Implementation Strategies), servant « à exprimer des informations sur les documents numériques en vue de leur pérennisation »<sup>389</sup>, tous deux utilisés par la BnF par exemple. Ces métadonnées « permettent d’assurer, dans le futur, l’intelligibilité des données et la mémoire du contexte exact dans lequel elles ont été collectées, ainsi que leurs conditions de restitution »<sup>390</sup>.

## Ontologies

Enfin, des outils de plus haut niveau permettent d’organiser l’information, de décrire les domaines de la connaissance, dans le web sémantique. Ces ontologies sont indispensables pour que l’interopérabilité puisse être effective : en effet, « pour qu’une autre application puisse décoder mon assertion, il faut que je donne une définition précise des identifiants que j’ai utilisés, qui sont sans doute différents de ceux que comprend cette application. En particulier, il faut que je donne une définition précise de mes “verbes” (“est situé à”) et mes “types” (Lieu, personne, etc.) »<sup>391</sup>. Ainsi, les ontologies ont pour objectif « de donner un sens univoque à ce dont je parle, à l’aide de la logique formelle (...). Les ontologies permettent également de déclarer des équivalences entre verbes ou entre types, rendant ainsi interopérables des données hétérogènes. Par exemple, je peux dire que, dans mon contexte, “est situé à” relie quelque chose à un “Lieu” et que cela représente la même notion que l’identifiant “basedNear” défini dans une autre ontologie bien connue, Foaf. Les ontologies font donc émerger de cet océan de liens des structures interopérables, rendant ainsi les données liées plus “sémantiques”, c’est-à-dire plus facilement réutilisables »<sup>392</sup>.

L’ontologie la plus fondamentale est **RDF**, qui modélise l’information sous forme de triplets interconnectés, où les éléments sont identifiés par une URI (Uniform Resource Identifier). RDF est le fondement même du web sémantique. Il met en relation des sujets et des objets, appartenant à des *classes* d’entités (par exemple Person, Book, Painting, Building, Event, PhilosophicalIdea...) au moyen de prédicats qui sont les *propriétés* s’appliquant à ces entités (par exemple createdBy, memberOf, successorTo, occurredAtTime, sameAs, familyName...). Ainsi, la propriété createdBy pourra mettre en relation une entité de la classe Person (comme objet) avec une autre entité de la classe Book (comme sujet)<sup>393</sup>.

RDF est un modèle très générique qui peut être réalisé dans différentes syntaxes, y compris XML (RDF-XML), mais aussi RDFa, Turtle, N-Triples, Json-LD... Les données en RDF d’une institution peuvent être stockées dans une base de données d’un type spécifique, nommée triplestore, interrogeable en Sparql.

Parmi les ontologies, les professionnels de l’IST seront particulièrement intéressés par **Skos**<sup>394</sup>, « une recommandation du W3C publiée le 18 août 2009 pour représenter des thésaurus documentaires, classifications ou d’autres types de vocabulaires contrôlés ou

<sup>388</sup> « Premis », Library of Congress, <http://www.loc.gov/standards/premis/>. Consulté le 9 février 2021

<sup>389</sup> « Premis : Preservation Metadata Implementation Strategies », BnF, <https://www.bnf.fr/fr/premis-preservation-metadata-implementation-strategies>. Consulté le 9 février 2021

<sup>390</sup> « Mets et Premis, outils pour les métadonnées de pérennisation », Doranum, <https://doranum.fr/stockage-archivage/mets-premis-outils-metadonnees-perennisation/>. Consulté le 9 février 2021

<sup>391</sup> Thomas Francart, « Le web de données, de « l’information en réseau » », *Arabesques* n° 83, 2016, <https://publications-prairial.fr/arabesques/index.php?id=522>. Consulté le 10 février 2021

<sup>392</sup> Ibid.

<sup>393</sup> Understanding Metadata : What is metadata, and what is it for? », 2017, Jenn Riley / NISO, p. 13, [https://groups.niso.org/apps/group\\_public/download.php/17446/Understanding%20Metadata.pdf](https://groups.niso.org/apps/group_public/download.php/17446/Understanding%20Metadata.pdf). Consulté le 1<sup>er</sup> février 2021

<sup>394</sup> « Skos Simple Knowledge Organization System », W3C, <https://www.w3.org/TR/skos-reference/>. Consulté le 10 février 2021

de langages documentaires »<sup>395</sup>, à l'aide d'attributs comme `skos:Concept`, `skos:prefLabel`, `skos:altLabel`, `skos:hiddenLabel`, `skos:broader` et `skos:narrower`, `skos:related`, etc., qui définissent la référence de prédicats dans des triplets RDF. Les vocabulaires Rameau et LCSH ont ainsi été publiés en Skos<sup>396</sup>.

Skos définit un ensemble de classes et de propriétés<sup>397</sup> permettant de décrire des thésaurus, c'est-à-dire une liste de termes d'un domaine du savoir et les relations sémantiques entre ces termes. Les données en Skos sont exprimées sous la forme de graphe, comme celui ci-dessous (exprimé en Turtle). Le graphe fait usage de la classe `skos:Concept` et de propriétés comme `skos:prefLabel` (label préférentiel), `skos:altLabel` (label alternatif), etc., pour exprimer des relations entre des concepts (amour, *synonyme* d'adoration ; émotion, concept *plus large* qu'amour)<sup>398</sup>, et formaliser un mini-thésaurus intitulé « My first thesaurus » (ce titre étant une propriété de Dublin Core, montrant comment les différents schémas de métadonnées peuvent s'articuler pour décrire une ressource, ici un thésaurus) :

```
<A> rdf:type skos:Concept ;
      skos:prefLabel "love"@en ;
      skos:altLabel "adoration"@en ;
      skos:broader <B> ;
      skos:inScheme <S> .

<B> rdf:type skos:Concept ;
      skos:prefLabel "emotion"@en ;
      skos:altLabel "feeling"@en ;
      skos:topConceptOf <S> .

<S> rdf:type skos:ConceptScheme ;
      dct:title "My First Thesaurus" ;
      skos:hasTopConcept <B>
```

Skos est lié au langage d'ontologie RDF **OWL** (Web Ontology Language), « conçu pour exprimer des structures conceptuelles complexes et riches (ontologies) supportant des fonctions logiques de contrôle de cohérence ou d'inférence »<sup>399</sup>. OWL est lui-même une extension de **RDF Schema** (RDFs), mais est plus expressif<sup>400</sup>.

**Schema.org** est une initiative de moteurs de recherche commerciaux et autres entreprises du numérique (lancée en 2011 par Google, Microsoft, Yahoo and Yandex) pour structurer les métadonnées d'un site web. Il appartient à la famille des microdonnées<sup>401</sup>. Il s'agit de données structurées en HTML encapsulant des informations

<sup>395</sup> « Simple Knowledge Organization System », Wikipédia, [https://fr.wikipedia.org/wiki/Simple\\_Knowledge\\_Organization\\_System](https://fr.wikipedia.org/wiki/Simple_Knowledge_Organization_System). Consulté le 10 février 2021

<sup>396</sup> « Rameau et Skos », Antoine Isaac et Thierry Bouchet, *Arabesques* n° 54, 2009, <https://publications-prairial.fr/arabesques/index.php?id=2109>. Consulté le 10 février 2021

<sup>397</sup> « Skos Simple Knowledge Organization System », W3C, <https://www.w3.org/TR/skos-reference/>. Consulté le 10 février 2021

<sup>398</sup> Ibid.

<sup>399</sup> « Simple Knowledge Organization System », Wikipédia, [https://fr.wikipedia.org/wiki/Simple\\_Knowledge\\_Organization\\_System](https://fr.wikipedia.org/wiki/Simple_Knowledge_Organization_System). Consulté le 10 février 2021

<sup>400</sup> « Web Ontology Language », Wikipédia, [https://fr.wikipedia.org/wiki/Web\\_Ontology\\_Language](https://fr.wikipedia.org/wiki/Web_Ontology_Language). Consulté le 10 février 2021

<sup>401</sup> « Microformat », Wikipédia, <https://fr.wikipedia.org/wiki/Microformat>. Consulté le 9 février 2021

lisibles par les moteurs de recherche<sup>402</sup>. Il propose un vocabulaire hiérarchisé constitué de 778 types et 1383 propriétés<sup>403</sup>. Les éléments permettant de décrire les « travaux créatifs » (creative works) sont au nombre de plusieurs dizaines<sup>404</sup>. Il revendique d'être utilisé par plus de 10 millions de sites<sup>405</sup>. Dans le monde des bibliothèques, Schema.org fait par exemple l'objet d'un développement par OCLC : « OCLC a fait le choix d'exposer les données de WorldCat en RDF avec le système de balisage Schema.org qui est utilisé conjointement par Google, Bing, Yahoo et Yandex, ce qui est un facteur de visibilité considérable. Le département Recherche d'OCLC participe activement au groupe Schema Bib Extend du W3C en vue de développer Schema.org pour la description bibliographique. Sans être très proches, BibFrame et Schema.org collaborent afin d'assurer un certain degré d'interopérabilité entre eux »<sup>406</sup>. Néanmoins, la page du Schema Bib Extend Working Group témoigne d'une activité mesurée ces dernières années<sup>407</sup>. Pour autant, l'utilisation de ce standard pour décrire des données bibliographiques est une possibilité envisagée par un certain nombre de professionnels<sup>408</sup>, et OCLC a développé l'extension de Schema.org, bib.schema.org, définissant « des microdonnées structurées pour différentes catégories de ressources (livres audio, thèse, etc.) »<sup>409</sup>.

## Vocabulaires contrôlés

L'indexation des ressources documentaires doit s'appuyer sur des termes non ambigus, notamment pour identifier les sujets et les thèmes des ressources indexées. Pour cela, on utilise des **vocabulaires contrôlés**.

En identifiant sans ambiguïté les concepts, un **répertoire de vedettes-matières** permet de décrire le contenu conceptuel des documents. Il s'agit d'« un lexique dont le but est de rendre possible l'organisation des connaissances afin d'optimiser la recherche d'information »<sup>410</sup>. Ils s'appuient sur la prédéfinition d'un ensemble de concepts permettant d'indexer les documents selon leurs sujets sans ambiguïté : « Les vocabulaires contrôlés permettent de résoudre les problèmes liés à l'homographie, la polysémie et la synonymie, par une relation bijective entre les concepts et les termes acceptés. »<sup>411</sup> Ils s'opposent en cela à l'indexation par mots clés libres.

Les **thésaurus** permettent plus spécifiquement d'établir des relations sémantiques (« relations de synonymie (terme équivalent), de hiérarchie (terme générique et terme

<sup>402</sup> « Taking Discoverability to the next Level: Datasets with DataCite DOIs Can Now Be Found through Google Dataset Search », Helena Cousijn et Martin Fenner, DataCite Blog, 2018, <https://blog.datacite.org/taking-discoverability-to-the-next-level/>. Consulté le 24 février 2021

<sup>403</sup> « Organization of schemas », Schema.org, <https://schema.org/docs/schemas.html>. Consulté le 9 février 2021

<sup>404</sup> « CreativeWork », Schema.org, <https://schema.org/CreativeWork>. Consulté le 9 février 2021

<sup>405</sup> Page d'accueil, Schema.org, <https://schema.org/>. Consulté le 9 février 2021

<sup>406</sup> « FRBR, RDA, Bibframe : comment prendre en compte ces nouveaux standards ? », Thierry Clavel, <https://books.openedition.org/pressesensib/6742#bodyftn17>. Consulté le 9 février 2021

<sup>407</sup> « Schema bib extend community group », W3C, <https://www.w3.org/community/schemabibex/>. Consulté le 9 février 2021

<sup>408</sup> « Multi-entity models of resource description in the Semantic Web: A comparison of FRBR, RDA and Bibframe », <https://www.emerald.com/insight/content/doi/10.1108/LHT-08-2014-0081/full/html> ; « Describing Theses and Dissertations Using Schema.org », Mixer, Jeffrey K., Patrick O'Brien, and Kenning Arlitsch. 2014, In DC-2014: Proceedings of the International Conference on Dublin Core and Metadata Applications. Austin, Texas, USA: Dublin Core Metadata Initiative, <https://www.oclc.org/research/publications/2014/describing-theses-using-schema.html> ; « A Division of Labor: The Role of Schema.org in a Semantic Web Model of Library Resources », Godby, Carol Jean, 2016, in « Linked Data for Cultural Heritage (Alets Monograph) », Ed Jones (ed.), Chicago: Amer Library Assn Editions », <https://www.oclc.org/research/publications/2016/schema-role-semantic-web-library.html>. Consultés le 9 février 2021

<sup>409</sup> Véronique Mesguich, *Bibliothèques : le web est à vous*, Electre-Cercle de la librairie, 2017, p. 82

<sup>410</sup> « Vocabulaire contrôlé », Wikipédia, [https://fr.wikipedia.org/wiki/Vocabulaire\\_contrôlé](https://fr.wikipedia.org/wiki/Vocabulaire_contrôlé). Consulté le 9 février 2021

<sup>411</sup> Ibid.

spécifique) et d'association (terme associé) »<sup>412</sup>) entre des termes représentant des concepts.

Parmi ces vocabulaires contrôlés, on compte notamment Rameau, géré le centre national Rameau, hébergé par la BnF, qui sert largement à l'indexation dans toutes les bibliothèques françaises. À l'international, les LCSH de la Library of Congress, ou encore pour la médecine MeSH et sa version française fMeSH<sup>413</sup>, sont des vocabulaires très importants.

Au-delà de ces vocabulaires génériques comme LCSH ou Rameau, ou portant sur des domaines très larges comme MeSH, il en existe de beaucoup plus spécialisés : on peut citer Animal Diseases Ontology, de l'Inrae<sup>414</sup>, et de nombreux autres vocabulaires à la granularité très fine fournis pour l'Inrae<sup>415</sup>. Car si les vocabulaires contrôlés ont vocation à être standards pour être partagés, leur champ d'application peut tout à fait être restreint à un seul champ disciplinaire. Cela introduit l'idée que la standardisation n'est pas forcément un appauvrissement des métadonnées, et qu'une multiplication de standards bien articulés entre eux permet aussi de rendre compte de la richesse de la production scientifique avec une granularité fine.

Cependant, pour qu'un vocabulaire contrôlé soit adopté et se répande, il faut qu'il réponde à certains critères, donc qu'il soit d'une relative simplicité d'utilisation. Par exemple, comme le note Étienne Cavalié, de la BnF, à propos de Rameau, dans l'introduction de *L'Indexation matière en transition*, pour expliquer la genèse de la réforme en cours de ce vocabulaire : « Sa structuration, le principe de vedettes pré-construites, les conditions d'utilisation (certains concepts ne peuvent s'appliquer qu'aux guerres, ou aux maladies), les règles de construction des chaînes d'indexation – bref, le manuel d'utilisation de Rameau de plusieurs centaines de pages, est un obstacle à une plus large adoption. Pourtant sa vocation à proposer un vocabulaire encyclopédique contrôlé pourrait satisfaire bon nombre de besoins, dans des domaines parfois très éloignés des bibliothèques ; et les technologies du web de données faciliteraient une utilisation complémentaire en ayant recours à des vocabulaires plus spécialisés. Mais la syntaxe Rameau est trop dépendante de strates historiques, ou de règles d'usages par thématiques. Elle n'est pas assez prédictive. Une ambition de diffusion plus large de ce vocabulaire doit donc passer préalablement par une réforme pour harmoniser les règles qui l'organisent, et en réduire considérablement le nombre. »<sup>416</sup>

En outre, s'ils facilitent l'interopérabilité, les vocabulaires contrôlés souffrent de plusieurs défauts : ils sont coûteux à créer et à maintenir, ils sont complexes et donc peu accessibles aux utilisateurs finaux, ils sont lents à mettre à jour (à l'inverse des moteurs de recherche reposant sur l'indexation automatisée des contenus) et ils sont éminemment subjectifs, ce qui rend encore plus difficile de s'entendre sur une version commune<sup>417</sup>.

Ainsi, par exemple, même dans une base de données utilisant un vocabulaire contrôlé, les utilisateurs peuvent préférer la recherche par mots clés dans le texte entier,

<sup>412</sup> Ibid.

<sup>413</sup> « Tout savoir sur le MeSH... ou presque », version d'avril 2010, Françoise Dailland, [http://documentation.abes.fr/sudoc/autres/2010Tout\\_savoir\\_sur\\_le\\_MeSH.pdf](http://documentation.abes.fr/sudoc/autres/2010Tout_savoir_sur_le_MeSH.pdf).

<sup>414</sup> « Animal Diseases Ontology », Inrae, <https://data.inrae.fr/dataset.xhtml?persistentId=doi:10.15454/1.44525654526207E12>. Exemple de vocabulaire FAIR cité par « Les principes FAIR », Doranum, <https://doranum.fr/enjeux-benefices/principes-fair/>. Consultés le 10 février 2021

<sup>415</sup> « Vocabulaires Ouverts @ INRAE », Inrae, <https://data.inrae.fr/dataverse/lovinra>. Consulté le 10 février 2021

<sup>416</sup> Étienne Cavalié, « Introduction », in Étienne Cavalié. *L'Indexation matière en transition - De la réforme de Rameau à l'indexation automatique*, Editions du Cercle de la Librairie, 2019 <https://hal.archives-ouvertes.fr/hal-02435623/file/L%27indexation%20matière%20en%20transition%20-%20Introduction%20-%20Etienne%20Cavalié.pdf>. Consulté le 10 février 2021

<sup>417</sup> « Introduction aux humanités numériques », op. cit. 3.4.7

s'il est disponible, ce qui engendre davantage de bruit et de silence mais qui est aussi plus immédiat d'utilisation.

## Protocoles d'échange de métadonnées

L'interopérabilité requiert des protocoles communément reconnus d'échanges de données entre bases.

Le protocole **Z39.50** est depuis 1988 « un protocole de recherche et d'échange d'informations bibliographiques en ligne, considéré comme pionnier des protocoles pour l'interrogation à distance »<sup>418</sup>. Il s'agit « d'un protocole de communication informatique client-serveur pour rechercher à travers un réseau informatique des informations dans des bases de données ». Il permet notamment d'interroger de façon simultanée des bases hétérogènes et de les afficher en temps réel. Il nécessite de disposer d'un client Z39.50, souvent fourni par les SIGB. Ainsi, par exemple, le Sudoc est interrogeable par Z39.50<sup>419</sup>, et les données de la BnF en Marc sont récupérables via ce protocole<sup>420</sup>. Il permet notamment aux bibliothèques de récupérer auprès d'autres acteurs des notices bibliographiques, et de gagner un temps précieux sur le travail de catalogage. Ainsi, les bibliothèques peuvent récupérer des notices auprès de prestataires de services comme Electre, ORB, Adav, etc., ou encore auprès de la BnF ou du Sudoc.

**SRU** (Search/Retrieve via URL) est l'adaptation du Z39.50 aux standards du web. Il permet de générer un webservice de recherche dans un catalogue de bibliothèques avec le langage Contextual Query Language (CQL)<sup>421</sup>. Un exemple est l'API<sup>422</sup> SRU BnF Catalogue général<sup>423</sup>.

L'Open Archives Initiative Protocol for Metadata Harvesting (**OAI-PMH**) est depuis le début des années 2000 un protocole permettant d'échanger des données entre bases, notamment des archives ouvertes et les entrepôts de données. Il permet de mettre à jour automatiquement des bases de données qui moissonnent d'autres bases ou archives pour en centraliser les données. L'institution moissonnée doit mettre à disposition ses données dans un entrepôt OAI qui permet de répondre aux requêtes des moissonneurs sous la forme de fichiers XML au format Dublin Core (au minimum).

Par exemple, BASE (Bielefeld Academix Search Engine), une base de données sans plein-texte, moissonne via OAI-PMH les métadonnées de plus de 8 000 autres bases de données, dont HAL et Theses.fr, les rendant ainsi plus visibles<sup>424</sup>. En revanche, si HAL est moissonné, il ne moissonne pas. Comme le notait déjà en 2014 Serge Bauin, dans « L'open access à moyen terme : une feuille de route pour HAL » : « Il n'est bien évidemment pas envisageable de connecter HAL avec toutes les sources possibles (...). Mais ne pourrait-on pas en identifier quelques-unes qui soient majeures, à la manière d'Isidore ? La question de "HAL moissonneur" se pose donc ici. Curieusement, l'idée

<sup>418</sup> « Comprendre les protocoles d'échange à travers deux exemples : Z39-50 et OAI-PMH », Charlotte Vignet, 2014, <https://www.enssib.fr/bibliotheque-numerique/documents/66636-comprendre-les-protocoles-d-echange-a-travers-deux-exemples-z39-50-et-oai-pmh.pdf>. Consulté le 10 février 2021

<sup>419</sup> « Interroger le Sudoc par le protocole Z39.50 », Abes, <http://www.abes.fr/Sudoc/Boite-a-outils-Sudoc-public/Z39.50-Sudoc-public>. Consulté le 10 février 2021

<sup>420</sup> « Récupérer les données de la BnF (au format Marc) », BnF, <https://www.bnf.fr/fr/recuperer-les-donnees-de-la-bnf-au-format-marc>. Consulté le 10 février 2021

<sup>421</sup> « The Contextual Query Language », Library of Congress, <https://www.loc.gov/standards/sru/cql/>. Consulté le 10 février 2021

<sup>422</sup> Une API (Application Programming Interface) est une interface qui rend disponible les données entre deux applications, via des requêtes selon des protocoles définis.

<sup>423</sup> « API SRU Catalogue général », BnF, <https://api.bnf.fr/fr/api-sru-catalogue-general>. Consulté le 10 février 2021

<sup>424</sup> « Content providers: By date », BASE, [https://www.base-search.net/about/en/about\\_sources\\_date.php](https://www.base-search.net/about/en/about_sources_date.php). Consulté le 10 février 2021

que HAL puisse moissonner est presque un tabou. Sans doute s'agit-il de s'assurer de la qualité : aucune source externe ne peut être conforme à tous les référentiels de métadonnées de HAL ! En un sens, le moissonnage est impossible. Pour poursuivre la métaphore agricole, le moissonnage collecterait trop de déchets, un tri manuel resterait nécessaire. »<sup>425</sup>

La BnF quant à elle utilise OAI-PMH à la fois pour moissonner et être moissonnée : « Le protocole OAI-PMH permet un double signalement des collections :

- le signalement à l'intérieur de Gallica de collections de bibliothèques partenaires (CNUM, Medica, Library of Congress, etc.) ;
- le signalement des documents numérisés de la BnF à l'extérieur de Gallica via d'autres catalogues, d'autres bibliothèques numériques, notamment Europeana. »<sup>426</sup>

L'information des bases de données peut également être récupérée sous la forme de triplets RDF dans des bases de données spécifiques nommées **triplestores**, interrogeables via un langage informatique propre, Sparql.

**Sword** (Simple Web-service Offering Repository Deposit) est une norme d'interopérabilité qui permet à des entrepôts numériques d'accepter le dépôt de contenus provenant de sources multiples dans différents formats (par exemple des documents XML) par l'intermédiaire d'un protocole standardisé. Contrairement à OAI-PMH, il s'agit d'un protocole d'interopérabilité pour le dépôt de données et non pour le moissonnage. HAL, qui ne moissonne pas, est en revanche compatible avec Sword : « L'API de dépôt Sword permet l'import automatique de documents dans l'archive ouverte HAL »<sup>427</sup>. Les entrepôts ArXiv, Dataverse, Dspace, Eprints, Fedora, Intralibrary, Microsoft Zenty, MyCoRe sont également compatibles<sup>428</sup>.

**KBart** (« Knowledge Bases and Related Tools ») est une recommandation de Niso<sup>429</sup> conçue pour faciliter le transfert de métadonnées entre les fournisseurs de contenus électroniques (éditeurs...) et les bases de connaissances, permettant d'améliorer la découverte des ressources dans un catalogue de bibliothèques, et leur consultation via OpenURL<sup>430</sup>. Une base de connaissances permet de recueillir l'information pour déterminer le droit à accéder à une ressource pour le lecteur d'une bibliothèque donnée (selon les abonnements souscrits par cette bibliothèque), et de le diriger vers la bonne version de cette ressource (celle à laquelle il a effectivement le droit d'accéder) grâce à un résolveur de liens. Mais des problèmes ont pu être identifiés dans la chaîne d'approvisionnement des métadonnées depuis les éditeurs vers les bases de connaissances, menant aux recommandations KBart en 2010, définissant les bonnes pratiques pour que les fournisseurs de contenus apportent des métadonnées de qualité<sup>431</sup>. En effet : « *Most metadata, like most content, originates from a publisher. In many cases, the metadata that is transferred in subsequent steps begins with the publisher, so if it is incorrect at the start, it will remain incorrect for most, if not all, of the remainder of the*

<sup>425</sup> « L'open access à moyen terme : une feuille de route pour HAL », Serge Bauin, <https://www.enssib.fr/bibliotheque-numerique/notices/64775-l-open-access-a-moyen-terme-une-feuille-de-route-pour-hal-hyper-articles-en-ligne>. Consulté le 10 février 2021

<sup>426</sup> « Les entrepôts OAI de la BnF », BnF, <https://www.bnf.fr/fr/les-entrepots-oai-de-la-bnf>. Consulté le 10 février 2020

<sup>427</sup> « Import Sword », HAL, <http://api.archives-ouvertes.fr/docs/sword>. Consulté le 10 février 2021

<sup>428</sup> « Sword (protocole) », Wikipédia, [https://fr.wikipedia.org/wiki/SWORD\\_\(protocole\)](https://fr.wikipedia.org/wiki/SWORD_(protocole)). Consulté le 10 février 2021

<sup>429</sup> « KBart: Knowledge Bases and Related Tools », NISO/UKSG KBART Working Group, 2010, <https://web.archive.org/web/20110716181454/http://www.niso.org/publications/rp/RP-2010-09.pdf>. Consulté le 10 février 2021

<sup>430</sup> « KBart Frequently Asked Questions », Niso, <https://www.niso.org/standards-committees/kbart/kbart-frequently-asked-questions>. Consulté le 10 février 2021

<sup>431</sup> « OpenURL knowledge base », Wikipédia, [https://en.wikipedia.org/wiki/OpenURL\\_knowledge\\_base](https://en.wikipedia.org/wiki/OpenURL_knowledge_base). Consulté le 10 février 2021



*supply chain.* »<sup>432</sup>

En 2014, une phase II de KBart élargit les recommandations à l'Open Access, et aux e-books : « *The importance of metadata for e-books has been apparent to the publishing, library, and e-commerce communities for some time. Bibliographic metadata relating to a specific book is still key and bibliographic metadata standards have evolved to take account of e-book formats.* »<sup>433</sup>

En 2019, Niso a publié KBart Automation, une nouvelle recommandation pour faciliter le transfert automatique de métadonnées via une API<sup>434</sup>.

À propos de KBart et des données des éditeurs, Yann Nicolas, de l'Abes, note que le format KBart rend possible des initiatives d'amélioration des données des éditeurs : « Les bibliothèques (ou d'autres acteurs) ont alors intérêt à se substituer aux éditeurs en valorisant et en libérant les métadonnées de leurs... fournisseurs. Cela suppose un effort de coordination mondiale, qui existe entre bibliothèques nationales, mais qui n'existe pas pour porter sur le web l'information scientifique et technique. Ce qui y ressemble le plus est néanmoins l'initiative GOKb (The Global Open Knowledgebase). Le périmètre de GOKb concerne les métadonnées de packages commerciaux. Il ne s'agit pas de décrire une revue électronique, mais le bouquet commercial qui donne accès à telle revue sur telle plateforme pour telle période de la vie de la revue (états de collection). GOKb s'est donné pour ambition de coordonner au niveau global la description sommaire de ces produits commerciaux, sous la forme technique du standard KBart. Il s'agit là de se substituer aux éditeurs pour décrire le contenu de leurs propres produits. Comme si vous deviez cataloguer votre Caddie... Cette situation semble absurde. Mais elle confirme la nécessité de prendre des initiatives qui, au mieux, viennent compléter et prolonger les efforts des éditeurs. »<sup>435</sup>

On peut encore mentionner **Onix** (Online Information eXchange)<sup>436</sup>, standard XML permettant d'échanger des métadonnées sur les livres et e-books, plutôt utilisé dans le domaine commercial que dans les bibliothèques. Néanmoins, comme le souligne l'Ifla<sup>437</sup>, ce format peut être utilisé par les agences bibliographiques, dans la mesure où il se substitue à une multitude de formats utilisés par les éditeurs, et qu'il peut être aligné sur Marc : « *Libraries have long been interested in the potential for using publisher information as a basis for catalogue records in order to improve efficiency. Publisher migration from proprietary local formats to Onix has made this a more realistic proposition by reducing the overhead in maintaining multiple translations to Marc.* » La BnF notamment récupère des éditeurs (ou éventuellement des distributeurs) des données sous format Onix pour le dépôt légal dématérialisé<sup>438</sup>. Cela facilite la démarche des éditeurs et permet à la BnF « de disposer de pré-notices avant le catalogage assuré dans les services de la bibliographie nationale »<sup>439</sup>. Néanmoins, le format Onix, du fait de sa

<sup>432</sup> « KBart: Knowledge Bases and Related Tools », Op. cit.

<sup>433</sup> « Knowledge Bases and Related Tools (KBart) Recommended Practice », NISO/ KBART Phase II Working Group, 2024, [https://groups.niso.org/apps/group\\_public/download.php/16900/RP-9-2014\\_KBART.pdf](https://groups.niso.org/apps/group_public/download.php/16900/RP-9-2014_KBART.pdf). Consulté le 10 février 2021

<sup>434</sup> « KBart Automation: Automated Retrieval of Customer Electronic Holdings », Niso/ KBart Automation Working Group, 2019, [https://groups.niso.org/apps/group\\_public/download.php/21896/NISO\\_RP-26-2019\\_KBART\\_Automation.pdf](https://groups.niso.org/apps/group_public/download.php/21896/NISO_RP-26-2019_KBART_Automation.pdf). Consulté le 10 février 2021

<sup>435</sup> « Atomes crochus. Les métadonnées des éditeurs et l'Abes », Yann Nicolas, in *Vers de nouveaux catalogues*, sous la direction d'Emmanuelle Bermès (2016)

<sup>436</sup> « Onix (publishing protocol) », Wikipédia, [https://en.wikipedia.org/wiki/ONIX\\_\(publishing\\_protocol\)](https://en.wikipedia.org/wiki/ONIX_(publishing_protocol)). Consulté le 10 février 2021

<sup>437</sup> « Onix (Online Information eXchange) », Ifla, <https://www.ifla.org/best-practice-for-national-bibliographic-agencies-in-a-digital-age/node/8859>. Consulté le 10 février 2021

<sup>438</sup> Voir notamment Jean-Charles Pajou, « Flux de données entre éditeurs et bibliothèques : le format Onix », in *Vers de nouveaux catalogues*, sous la direction d'Emmanuelle Bermès, Op. Cit.

<sup>439</sup> Ibid.

vocation commerciale de départ, ne comprend pas nécessairement, ou de façon inégale, toutes les métadonnées nécessaires à un catalogage de qualité : « En règle générale, les données ne sont pas aussi riches ou précises que les principes qui régissent les normes ou les pratiques de catalogage et d'indexation en bibliothèque. »<sup>440</sup>

En outre, le standard Onix for Publications Licenses (Onix-PL) permet de gérer les autorisations d'utilisation des ressources numériques notamment par les bibliothèques<sup>441</sup>.

De façon un peu différente, puisqu'il s'agit moins d'échange que de récupération de métadonnées, **Zotero** est un outil de gestion bibliographique intégré aux navigateurs web depuis son lancement en 2006 sur Mozilla Firefox. Il propose des fonctionnalités impressionnantes de capture des métadonnées ainsi que des fichiers plein-texte lorsqu'ils sont disponibles. Les utilisateurs peuvent ensuite éditer et organiser ces références selon leurs besoins.

Zotero est un éloquent exemple de la puissance que peuvent avoir les métadonnées, si des outils performants et de préférence libres sont mis à disposition pour les exploiter : « *Zotero can automatically add publication data by DOI or ISBN and find open-access PDFs when you don't have access to a paper. You can create advanced searches — say, all articles mentioning a certain keyword added in the last month — and save them as auto-updating collections. When you open a paywalled page in your browser, Zotero can automatically redirect you through your institution's proxy so that you can access the PDF. Zotero can even warn you if you try to cite a paper that was retracted.* »<sup>442</sup>

Pour résumer, les acteurs du monde des ressources scientifiques communiquent leurs données, et le font à travers une multiplicité de moyens, adaptés aux humains et/ou aux machines, permettant de récolter des données en masse ou en petite quantité, sous différents formats, pour différentes catégories d'utilisateurs. À titre d'exemple, voici ci-dessous un tableau élaboré par Crossref<sup>443</sup> pour résumer les moyens d'échanges de données qu'il propose, ainsi que leurs principales caractéristiques :

<sup>440</sup> Ibid.

<sup>441</sup> « Onix-PL », Onix, <https://www.editeur.org/21/ONIX-PL/>. Consulté le 10 février 2021

<sup>442</sup> « Why Zotero », Zotero, <https://www.zotero.org/why> ; pour un exemple d'enregistrement automatique du statut (ici : rétracté) des articles : tweet de M. Sierra-Arévalo du 14 juillet 2020, <https://twitter.com/michaelsierraa/status/1283098707307749378?s=20>. Consultés le 10 février 2021

<sup>443</sup> « Metadata retrieval », Crossref, <https://www.crossref.org/education/retrieve-metadata/>. Consulté le 10 février 2021

Feature / option	Metadata Search	Simple Text Query	REST API	XML API	OAI-PMH	OpenURL	Metadata Plus (OAI-PMH + REST API)
Interface for people or machines?	People	People	People (low volume and occasional use) and machines	Machines	Machines	Machines	Machines
Output format	Text, JSON	Text	JSON	XML	XML	XML	JSON, XML
Suitable for citation matching?	Yes (low volume)	Yes	Yes	Yes	No	No	Yes
Supports volume downloads?	No	No	Yes	No	Yes	No	Yes
Suitable for usage type	Frequent and occasional	Frequent and occasional	Frequent and occasional	Frequent	Frequent	Frequent	Frequent and occasional
Free or cost?	Free	Free	Free and cost options	Free and cost options	Cost for full service, more options available	Free but requires an account	Cost
Includes all available metadata?	In JSON only	DOIs only	Yes	Yes	Yes	Bibliographic only	Yes
Documentation	<a href="#">Metadata Search</a>	<a href="#">Simple Text Query</a>	<a href="#">REST API</a>	<a href="#">XML API</a>	<a href="#">OAI-PMH</a>	<a href="#">OpenURL</a>	<a href="#">Metadata Plus (OAI-PMH + REST API)</a>

### Propriétés des protocoles de récupération de données depuis Crossref

## BIBLIOGRAPHIE

---

AGENCE BIBLIOGRAPHIQUE DE L'ENSEIGNEMENT SUPÉRIEUR, 2014. « Métarevues » : un outil dédié au traitement des périodiques. In : *PUNKTOKOMO* [en ligne]. [Consulté le 11 janvier 2021]. Disponible à l'adresse : <https://punktokomo.abes.fr/2014/06/26/metarevues-un-outil-dedie-au-traitement-des-periodiques/>.

AGENCE BIBLIOGRAPHIQUE DE L'ENSEIGNEMENT SUPÉRIEUR, 2017. Autorités vs référentiels : 3 questions aux experts de l'Abes. In : *PUNKTOKOMO* [en ligne]. [Consulté le 25 mai 2020]. Disponible à l'adresse : <https://punktokomo.abes.fr/2017/04/20/autorites-vs-referentiels-3-questions-aux-experts-de-labes/>.

AGENCE BIBLIOGRAPHIQUE DE L'ENSEIGNEMENT SUPÉRIEUR, 2018. Projet d'établissement 2018-2022 [en ligne]. [Consulté le 28 février 2021]. Disponible à l'adresse : <https://abes.fr/publications/publications-institutionnelles/projet-etablissement-2018-2022/>

AGENCE BIBLIOGRAPHIQUE DE L'ENSEIGNEMENT SUPÉRIEUR, 2020. L'alignement des identifiants auteurs entre IdRef & HAL : un état des lieux. In : *PUNKTOKOMO* [en ligne]. [Consulté le 2 octobre 2020]. Disponible à l'adresse : <https://punktokomo.abes.fr/2020/10/02/lalignement-des-identifiants-auteurs-entre-idref-hal-un-etat-des-lieux/>.

AGENCE BIBLIOGRAPHIQUE DE L'ENSEIGNEMENT SUPÉRIEUR, 2021. Le signalement des corpus acquis sous licence nationale évolue ! In : *PUNKTOKOMO* [en ligne]. [Consulté le 29 janvier 2021]. Disponible à l'adresse : <https://punktokomo.abes.fr/2021/01/19/le-signalement-des-corpus-acquis-en-licence-nationale-evolue/>.

AGENCE BIBLIOGRAPHIQUE DE L'ENSEIGNEMENT SUPÉRIEUR, 2013. *Étude sur la faisabilité et le positionnement d'un hub de métadonnées ABES* [en ligne]. [Consulté le 4 janvier 2021]. Disponible à l'adresse : [https://abes.fr/wp-content/uploads/2020/02/Hub\\_versionFinale\\_5juillet2013.pdf](https://abes.fr/wp-content/uploads/2020/02/Hub_versionFinale_5juillet2013.pdf).

AKERROYD, John, 2017. Discovery systems: Are they now the library? In : *Learned Publishing*. 2017. Vol. 30, n° 1, p. 87-89. DOI 10.1002/leap.1085. Disponible à l'adresse : <https://onlinelibrary.wiley.com/doi/full/10.1002/leap.1085>

ALARCON, Nicolas et DELHAYE, Marlène, 2019. CasuHAL au service de la science ouverte. In : *Arabesques* n° 93 [en ligne]. [Consulté le 6 avril 2020]. Disponible à l'adresse : <https://publications-prairial.fr/arabesques/index.php?id=544>.

ALI, Mufazil, LOAN, Fayaz Ahmad et MUSHATQ, Rabiya, 2018. Open Access Scientific Digital Repositories : An Analytical Study of the OpenDOAR. In : *2018 5th International Symposium on Emerging Trends and Technologies in Libraries and Information Services (ETTLIS)* [en ligne]. Noida : IEEE. p. 213-216. [Consulté le 6 avril 2020]. ISBN 978-1-5386-0828-9. Disponible à l'adresse : <https://ieeexplore.ieee.org/document/8485265/>.

ALLISON-CASSIN, Stacy et SCOTT, Dan, 2018. Wikidata: a platform for your library's linked open data. In : *The Code4Lib Journal* n° 40 [en ligne]. [Consulté le 15 février 2021]. Disponible à l'adresse : <https://journal.code4lib.org/articles/13424>.

AMIEL, Philippe, FRONTINI, Francesca, LACOUR, Pierre-Yves et ROBIN, Agnès, 2020. Pratiques de gestion des données de la recherche : une nécessaire acculturation des chercheurs aux enjeux de la science ouverte ? Résultats d'une enquête exploratoire dans le bassin montpelliérain (juin 2018). In : *Cahiers Droit, Sciences & Technologies* n° 10, p. 147-168. DOI 10.4000/cdst.2061. Disponible à l'adresse : <https://journals.openedition.org/cdst/2061>.

ANADIOTIS, George, 2017. Graph databases and RDF: It's a family affair. In : *ZDNet* [en ligne]. [Consulté le 29 avril 2020]. Disponible à l'adresse : <https://www.zdnet.com/article/graph-databases-and-rdf-its-a-family-affair/>.

AZEROUAL, Otmane et LEWONIEWSKI, Włodzimierz, 2020. How to Inspect and Measure Data Quality about Scientific Publications: Use Case of Wikipedia and CRIS Databases. In : *Algorithms*. Vol. 13, n° 5. DOI 10.3390/a13050107. Disponible à l'adresse : [https://www.researchgate.net/publication/340947946\\_How\\_to\\_Inspect\\_and\\_Measure\\_Data\\_Quality\\_about\\_Scientific\\_Publications\\_Use\\_Case\\_of\\_Wikipedia\\_and\\_CRIS\\_Databases](https://www.researchgate.net/publication/340947946_How_to_Inspect_and_Measure_Data_Quality_about_Scientific_Publications_Use_Case_of_Wikipedia_and_CRIS_Databases).

BACA, Murtha, (dir.), 2016. *Introduction to Metadata*. 3rd ed., Los Angeles: Getty Publications [Consulté le 31 août 2020]. Disponible à l'adresse : <http://www.getty.edu/publications/intrometadata>.

BALIGAND, Marie-Pascale, COLCANAP, Grégory, HARNAIS, Vincent, ROUSSEAU-HANS, Françoise, WEIL-MIKO, Christine, 2021. Les pratiques de recherche documentaire des chercheurs français en 2020 : étude du consortium Couperin. Rapport Couperin n° 2, Couperin.org, (hal-03148285) [en ligne]. [Consulté le 2 mars 2021]. Disponible à l'adresse <https://hal.inrae.fr/hal-03148285/document>. Consulté le 2 mars 2021

BASCONES, Magaly et STANIFORTH, Amy, 2018. What is all this fuss about? Is wrong metadata *really* bad for libraries and their end-users? In : *Insights*. 24 octobre 2018. Vol. 31. DOI 10.1629/uksg.441. Disponible à l'adresse : <https://insights.uksg.org/articles/10.1629/uksg.441/>.

BAUIN, Serge, 2014. *L'open access à moyen terme : une feuille de route pour HAL* [en ligne]. [Consulté le 6 avril 2020]. Disponible à l'adresse : [http://corist-shs.cnrs.fr/sites/default/files/billets/cnrs\\_dist\\_rapport\\_bauin\\_sur\\_ccsd\\_et\\_hal\\_septembre\\_2014.pdf](http://corist-shs.cnrs.fr/sites/default/files/billets/cnrs_dist_rapport_bauin_sur_ccsd_et_hal_septembre_2014.pdf).

BÉQUET, Gaëlle et OURY, Clément, 2018. Revisiting the identification of serials: ISSN goes linked. In : *Insights*. 1 mars 2018. Vol. 31, n° 0, p. 2, DOI 10.1629/uksg.402. Disponible à l'adresse : <https://insights.uksg.org/articles/10.1629/uksg.402/>

BERMÈS, Emmanuelle, 2016 (dir.). *Vers de nouveaux catalogues*, Éditions du Cercle de la librairie, « Bibliothèques ». ISBN : 9782765415138. DOI : 10.3917/elec.berme.2016.01. Disponible à l'adresse : <https://www.cairn.info/vers-de-nouveaux-catalogues--9782765415138.htm>.

BERMÈS, Emmanuelle, 2020. *Le numérique en bibliothèque : naissance d'un patrimoine : l'exemple de la Bibliothèque nationale de France (1997-2019)* [en ligne]. Thèse de doctorat. Paris, École nationale des chartes. [Consulté le 31 août 2020]. Disponible à l'adresse : <https://tel.archives-ouvertes.fr/tel-02475991>.

BIBLIOTHÈQUE NATIONALE DE FRANCE, 2019. Qu'est-ce que les données d'autorité ? In : *BnF - Site institutionnel* [en ligne]. [Consulté le 6 juillet 2020]. Disponible à l'adresse : <https://www.bnf.fr/fr/mediatheque/quest-ce-que-les-donnees-dautorite>.

BIBLIOTHÈQUE NATIONALE DE FRANCE, 2021. Politique identifiants BnF. In : *BnF - Site institutionnel* [en ligne]. [Consulté le 26 janvier 2021]. Disponible à l'adresse : <https://www.bnf.fr/fr/politique-identifiants-bnf>.

BILDER G., LIN J., NEYLON C., 2020. The Principles of Open Scholarly Infrastructure. In : *The Principles of Open Scholarly Infrastructure* [en ligne]. [Consulté le 8 décembre 2020]. Disponible à l'adresse : <https://openscholarlyinfrastructure.org/>.

BIZOS, Isabelle, 2020. *Big deals et open access : quelle stratégie numérique pour les bibliothèques universitaires ?* Enssib [en ligne]. [Consulté le 6 novembre 2020]. Disponible à l'adresse : <https://www.enssib.fr/bibliotheque-numerique/notices/69602-big-deals-et-open-access-quelle-strategie-numerique-pour-les-bibliotheques-universitaires>.

BLANCHARD, Antoine et SABUNCU, Elifsu, 2015. *Pour une meilleure visibilité de la recherche française*, Deuxième labo [en ligne]. [Consulté le 29 décembre 2020]. Disponible à l'adresse : <https://hal.archives-ouvertes.fr/hal-01251541>.

BOUDRY, Christophe, 2021. Availability of ORCID in publications archived in PubMed, MEDLINE, and Web of Science Core Collection. In : *Scientometrics* [en ligne]. [Consulté le 16 février 2021]. DOI : 10.1007/s11192-020-03825-7. Disponible à l'adresse : <https://doi.org/10.1007/s11192-020-03825-7>.

BOUDRY, Christophe et DURAND-BARTHEZ, Manuel, 2020. Use of author identifier services (ORCID, ResearchID) and academic social networks (Academia.edu, ResearchGate) by the researchers of the University of Caen Normandy (France): A case study. In : *PLOS ONE*. Vol. 15, n° 9. DOI 10.1371/journal.pone.0238583. Disponible à l'adresse : <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0238583>

BOURDENET, Philippe, 2013. OCLC, l'histoire d'une coopération fructueuse. In : *Documentaliste - Sciences de l'information*. 2013. Vol. 50, p. 30.

BROWN, John, HAAK, Laurel L., MEADOWS, Alice, 2018. Using ORCID, DOI, and Other Open Identifiers in Research Evaluation | Research Metrics and Analytics. In : *Frontiers in Research Metrics and Analytics* 2018, n°3 [en ligne]. [Consulté le 9 avril 2020]. Disponible à l'adresse : <https://www.frontiersin.org/articles/10.3389/frma.2018.00028/full>.

BULL, Sarah et QUIMBY, Amanda, 2016. A renaissance in library metadata? The importance of community collaboration in a digital world. In : *Insights*, Vol. 29, n° 2, p. 146-153. DOI 10.1629/uksg.302. Disponible à l'adresse : <https://insights.uksg.org/articles/10.1629/uksg.302/>

BURNARD, Lou, 2015. *Qu'est-ce que la Text Encoding Initiative ?* [en ligne]. OpenEdition Press, Encyclopédie numérique. [Consulté le 27 février 2021]. ISBN 978-2-8218-5581-6. Disponible à l'adresse : <https://books.openedition.org/oep/1237>.

CAR, Nicholas, GOLODONIUC, Pavel et KLUMP, Jens, 2017. The Challenge of Ensuring Persistency of Identifier Systems in the World of Ever-Changing Technology. In : *Data Science Journal*. Vol. 16, n° 0. DOI 10.5334/dsj-2017-013.

CARACO, Alain, 2019. Open access et bibliothèques. In : *Arabesques* n° 93 [en ligne]. DOI 10.35562/arabesques.543. Disponible à l'adresse : <https://publications-prairial.fr/arabesques/index.php?id=543>.

CASUHAL, 2020. *Enquête Adhérents 2020 - Synthèse des résultats* [en ligne]. [Consulté le 25 septembre 2020]. Disponible à l'adresse : <https://www.casuhal.org/wp-content/uploads/sites/21/2020/09/Casuhal-Enqu%C3%AAt%20Adh%C3%A9rents-2020-Synth%C3%A8se-v1.pdf>.

CAVALIÉ, Étienne, 2018. Des milliers de ebooks (et de liens !) dans le catalogue de la BnF. In : *Bibliothèques [reloaded]* [en ligne]. [Consulté le 25 mai 2020]. Disponible à l'adresse : <https://bibliotheques.wordpress.com/2018/07/09/des-milliers-de-ebooks-et-de-liens-dans-le-catalogue-de-la-bnf/>.

CAVALIÉ, Étienne (dir.), 2019. *L'indexation matière en transition. De la réforme de Rameau à l'indexation automatique*, Éditions du Cercle de la librairie, Collection « Bibliothèques », ISBN 978-2-7654-1623-4

CAVALIÉ, Étienne, 2019. Les chantiers d'indexation rétrospective à la Bibliothèque nationale de France. In : CAVALIÉ, Étienne (dir.), 2019 [Consulté le 21 mai 2020]. Disponible à l'adresse : <https://hal.archives-ouvertes.fr/hal-02435624>.

CAVALIÉ, Étienne, 2019. Les chantiers d'indexation rétrospective à la Bibliothèque nationale de France. In : CAVALIÉ, Étienne (dir.), 2019 [Consulté le 21 mai 2020]. Disponible à l'adresse : <https://hal.archives-ouvertes.fr/hal-02435624>.

CAVALIÉ, Étienne, 2019. L'indexation matière en transition - De la réforme de Rameau à l'indexation automatique : Conclusion. In CAVALIÉ, Étienne (dir.),

2019 [Consulté le 21 mai 2020]. Disponible à l'adresse : <https://hal.archives-ouvertes.fr/hal-02435623>.

CCSD, 2016. Feuille de route du CCSD 2016-2020 [en ligne]. [Consulté le 25 septembre 2020]. Disponible à l'adresse : <https://www.ccsd.cnrs.fr/2016/06/feuille-de-route-du-ccsd-2016-2020/>.

CLARE, Helen, 2018. Open access briefing paper: The potential of global identifiers to support more efficient workflows for all kinds of OA. In : *Jisc scholarly communications* [en ligne]. [Consulté le 14 mai 2020]. Disponible à l'adresse : <https://scholarlycommunications.jiscinvolve.org/wp/2018/10/24/open-access-briefing-paper-the-potential-of-global-identifiers-to-support-more-efficient-workflows-for-all-kinds-of-oa/>.

CLAVEL, Thierry, 2019. FRBR, RDA, BibFrame : comment prendre en compte ces nouveaux standards ? In : SVENBRO, Anna (éd.), *Réinformatiser une bibliothèque* [en ligne]. Villeurbanne : Presses de l'enssib. La Boîte à outils. [Consulté le 9 octobre 2020]. ISBN 978-2-37546-093-1. Disponible à l'adresse : <http://books.openedition.org/pressesenssib/6742>.

COALITION S. « Plan S » and « cOAlition S » – Accelerating the transition to full and immediate Open Access to scientific publications [en ligne]. [Consulté le 29 septembre 2020]. Disponible à l'adresse : <https://www.coalition-s.org/>.

COMITÉ POUR LA SCIENCE OUVERTE (Collège Europe et international), 2019. Des identifiants ouverts pour la science ouverte : note d'orientation. In : *Ouvrir la science* [en ligne]. [Consulté le 3 avril 2020]. Disponible à l'adresse : <https://www.ouvrirlascience.fr/des-identifiants-ouverts-pour-la-science-ouverte-note-dorientation/>.

CONDITOR (équipe), 2020. CONDITOR – bilan [en ligne]. [Consulté le 16 octobre 2020]. Disponible à l'adresse : <https://www.ouvrirlascience.fr/conditor-3>.

COX, Andrew M., KENNAN, Mary Anne, LYON, Liz, PINFIELD, Stephen et SBAFFI, Laura, 2019. Maturing research data services and the transformation of academic libraries. In : *Journal of Documentation*. 26 septembre 2019. Vol. 75, n° 6, p. 1432-1462. DOI 10.1108/JD-12-2018-0211. Disponible à l'adresse : <https://www.emerald.com/insight/content/doi/10.1108/JD-12-2018-0211/full/html>

CUXAC, Pascal, COLLIGNON, Alain, GREGORIO, Stéphanie et PARMENTIER, François, 2019. Istex Des bases de données massives au Web de données : désambiguïsation et alignement d'entités géographiques dans les textes scientifiques. 12<sup>e</sup> Colloque international d'ISKO-France : Données et mégadonnées ouvertes en SHS : de nouveaux enjeux pour l'état et l'organisation des connaissances ? Octobre 2019, Montpellier, France. Disponible à l'adresse : <https://hal.archives-ouvertes.fr/hal-02307577/document>

CUXAC, Pascal et THOUVENIN, Nicolas, 2017. Archives numériques et fouille de textes : le projet ISTEEX. Disponible à l'adresse :



[https://www.researchgate.net/publication/312233362\\_Archives\\_numeriques\\_et\\_fo\\_uille\\_de\\_textes\\_le\\_projet\\_ISTEX](https://www.researchgate.net/publication/312233362_Archives_numeriques_et_fo_uille_de_textes_le_projet_ISTEX)

DA SYLVA, Lyne, 2017. Les données et leurs impacts théoriques et pratiques sur les professionnels de l'information. In : *Documentation et bibliothèques*, vol. 63, n° 4, p. 5-34. DOI 10.7202/1042308ar. Disponible à l'adresse : <https://www.erudit.org/fr/revues/documentation/2017-v63-n4-documentation03290/1042308ar/>

DELEMONTÉZ, Renaud, 2017. *Les bibliothèques universitaires face aux systèmes d'information recherche : nouveaux outils, nouveaux rôles ?*, Enssib [en ligne]. [Consulté le 28 décembre 2020]. Disponible à l'adresse : <https://www.enssib.fr/bibliotheque-numerique/documents/67434-les-bibliotheques-universitaires-face-aux-systemes-d-information-recherche-nouveaux-outils-nouveaux-roles.pdf>.

DEMPSEY, Lorcan, 2016. Library collections in the life of the user: two directions. In : *LIBER Quarterly*, vol. 26, n° 4, p. 338-359. DOI 10.18352/lq.10170. Disponible à l'adresse : <https://www.liberquarterly.eu/articles/10.18352/lq.10170/>.

DE WILDE, Max, GILLET Florence, HENGCHEN, Simon, VAN HOOLAND, Seth, 2016. *Introduction aux humanités numériques : méthodes et pratiques*, De Boeck Supérieur, ISBN : 9782807302150.

DUDEK, Jonathan, MONGEON, Philippe et BERGMANS, Josephine, 2019. DataCite as a Potential Source for Open Data Indicators. In : 17th International Conference on Scientometrics & Informetrics, 2-5 septembre 2019. Disponible à l'adresse : <https://crtcscs.openum.ca/files/sites/60/2019/09/ISSI2019-datacite-potential-source-open-data-indicators.pdf>

EUROPEAN COMMISSION. DIRECTORATE GENERAL FOR RESEARCH AND INNOVATION. et EOSC EXECUTIVE BOARD, 2021. *Digital skills for FAIR and Open Science: report from the EOSC Executive Board Skills and Training Working Group* [en ligne]. LU : Publications Office. [Consulté le 22 février 2021]. Disponible à l'adresse : <https://data.europa.eu/doi/10.2777/59065>.

FENNER, Martin, 2020. DataCite Commons - Exploiting the Power of PIDs and the PID Graph. In : *DataCite Blog* [en ligne]. [Consulté le 29 août 2020]. Disponible à l'adresse : <https://blog.datacite.org/power-of-pids/>.

FENNER, Martin, 2020. Making the most out of available Metadata. In : *DataCite Blog* [en ligne]. [Consulté le 30 août 2020]. Disponible à l'adresse : <https://blog.datacite.org/making-the-most/>.

FRANCART, Thomas, 2016. Le web de données, de « l'information en réseau ». In : *Arabesques* n° 83 [en ligne]. [Consulté le 7 avril 2020]. Disponible à l'adresse : <https://publications-prairial.fr/arabesques/index.php?id=522>.

FREYA, 2018. *Survey of Current PID Services Landscape* [en ligne]. [Consulté le 3 avril 2020 by]. Disponible à l'adresse : [https://www.project-freya.eu/en/deliverables/freya\\_d3-1.pdf](https://www.project-freya.eu/en/deliverables/freya_d3-1.pdf).

- GÉROUDET, Marie-Madeleine, CROHEM, Laurence et MALLERET, Cécile, 2020. Un écosystème pour la visibilité des productions scientifiques : l'expérience de l'université de Lille. In : *Arabesques n° 95* [en ligne]. [Consulté le 31 décembre 2020]. Disponible à l'adresse : <https://publications-prairial.fr/arabesques/index.php?id=1303>.
- GODBY, Carol Jean, 2017. A Division of Labor: The Role of Schema.org in a Semantic Web Model of Library Resources. In : In *Linked Data for Cultural Heritage* (Alcts Monograph), Ed Jones (dir.). Chicago: Amer Library Assn Editions, p. 32. [Consulté le 26 février 2021]. Disponible à l'adresse : <https://www.oclc.org/content/dam/research/publications/2017/godby-division-of-labor-2017.pdf>
- GREGG, Will, ERDMANN, Christopher, PAGLIONE, Laura, SCHNEIDER, Juliane et DEAN, Clare, 2019. A literature review of scholarly communications metadata. In : *Research Ideas and Outcomes*, vol. 5, pp. e38698. DOI 10.3897/rio.5.e38698. [Consulté le 26 février 2021]. Disponible à l'adresse : <https://riojournal.com/article/38698/>
- GUÉDON, Jean-Claude, 2001. À l'ombre d'Oldenburg : bibliothécaires, chercheurs scientifiques, maisons d'édition et le contrôle des publications scientifiques. ARL Meeting, mai 2001, Toronto, Canada. [Consulté le 26 février 2021]. Disponible à l'adresse : <https://halshs.archives-ouvertes.fr/halshs-00395366/document>
- GUÉDON, Jean-Claude, 2019. *Future of scholarly publishing and scholarly communication: report of the expert group to the European Commission* [en ligne]. [Consulté le 25 août 2020]. ISBN 978-92-79-97238-6. DOI : <https://doi.org/10.2777/836532>. Disponible à l'adresse : <https://op.europa.eu/en/publication-detail/-/publication/464477b3-2559-11e9-8d04-01aa75ed71a1/language-en>
- HEIBI, Ivan, PERONI, Silvio et SHOTTON, David, 2019. Crowdsourcing open citations with Croci -An analysis of the current status of open citations, and a proposal [en ligne]. [Consulté le 9 avril 2020]. Disponible à l'adresse : <http://arxiv.org/abs/1902.02534>.
- HEIBI, Ivan, PERONI, Silvio et SHOTTON, David, 2019b. Software review: Coci, the OpenCitations Index of Crossref open DOI-to-DOI citations. In : *Scientometrics*, vol. 121, n° 2, pp. 1213-1228. DOI 10.1007/s11192-019-03217-6. Disponible à l'adresse : <https://link.springer.com/article/10.1007/s11192-019-03217-6>
- HENDRICKS, Ginny, TKACZYK, Dominika, LIN, Jennifer et FEENEY, Patricia, 2020. Crossref: The sustainable source of community-owned scholarly metadata. In : *Quantitative Science Studies*, vol. 1, n° 1, p. 414-427. DOI 10.1162/qss\_a\_00022. Disponible à l'adresse : [https://www.mitpressjournals.org/doi/full/10.1162/qss\\_a\\_00022?mobileUi=0](https://www.mitpressjournals.org/doi/full/10.1162/qss_a_00022?mobileUi=0)
- HEUSSE, Marie-Dominique, 2017. Faire parler les données de la recherche grâce au Web sémantique : le projet VIVO. In : *1er Atelier Valorisation et Analyse des Données de la Recherche (VADOR 2017) organisé durant la 35e édition du*

*congrès Informatique des Organisations et Systèmes d'Information et de Décision (INFORSID 2017)* [en ligne]. Toulouse, France, mai 2017. p. 19-25. [Consulté le 22 janvier 2021]. Disponible à l'adresse : <https://hal.archives-ouvertes.fr/hal-01887894>.

HO, Jeannette, 2021. The Roles of Cataloging vs. Non-Cataloging Librarians and Staff in Non-MARC Metadata-Related Workflows: A Survey of Academic Libraries in the United States. In : *Cataloging & Classification Quarterly*, vol. 50, p. 1-33. DOI 10.1080/01639374.2020.1863889.

ILLIEN, Gildas et BOURDON, Françoise, 2014. À la recherche du temps perdu, retour vers le futur : CBU 2.0 [en ligne]. Disponible à l'adresse : <http://library.ifla.org/956/1/086-illien-fr.pdf>.

IFLA, 2017. Bibliographic control [en ligne]. [Consulté le 25 mai 2020]. Disponible à l'adresse : <https://www.ifla.org/best-practice-for-national-bibliographic-agencies-in-a-digital-age/node/8911>.

IFLA WIKIDATA WORKING GROUP, 2020. Wikicite + Libraries Discussion Series - YouTube [en ligne]. [Consulté le 16 février 2021 b]. Disponible à l'adresse : <https://www.youtube.com/playlist?list=PLV81siTMahbsjakNDIdFwbIvdyNqWRN>  
[Po.](#)

INIST-CNRS, 2017. Identifiants pérennes : fiche synthétique. In : *DoRANum*. [en ligne]. [Consulté le 3 avril 2020]. Disponible à l'adresse : <https://doranum.fr/identifiants-perennes-pid/fiche-synthetique/>.

JACSO, Peter, 2005. As we may search - Comparison of major features of the Web of Science, Scopus, and Google Scholar citation-based and citation-enhanced databases. In : *Current science*, vol. 89, p. 1537-1547. Disponible à l'adresse : [https://www.researchgate.net/publication/234144770\\_As\\_we\\_may\\_search-Comparison\\_of\\_major\\_features\\_of\\_the\\_Web\\_of\\_Science\\_Scopus\\_and\\_Google\\_Scholar\\_citation-based\\_and\\_citation-enhanced\\_databases](https://www.researchgate.net/publication/234144770_As_we_may_search-Comparison_of_major_features_of_the_Web_of_Science_Scopus_and_Google_Scholar_citation-based_and_citation-enhanced_databases)

JEANGIRARD, Éric, 2019. Monitoring Open Access at a national level: French case study. In : *ELPUB 2019 23rd edition of the International Conference on Electronic Publishing* [en ligne]. Marseille, France. [Consulté le 31 août 2020]. Disponible à l'adresse : <https://hal.archives-ouvertes.fr/hal-02141819>.

JEULIN, Michael, LE PROVOST, Aline et OLIVIER, Yann, 2016. Linked data, enjeu(x) et devenir, in : *Arabesques* n° 83 [En ligne]. [Consulté le 14 juin 2020]. Disponible à l'adresse : <https://publications-prairial.fr/arabesques/index.php?id=651>.

JOST, Clémence, 2019. Les bibliothèques françaises dans Worldcat : l'Abes souhaite une concertation avec la Bnf et OCLC. In : *Archimag* [en ligne]. [Consulté le 15 mai 2020]. Disponible à l'adresse : <https://www.archimag.com/bibliotheque-edition/2019/03/19/bibliotheques-fran%C3%A7aises-worldcat-abes-concertation-bnf-oclc>.

KEMP, Jennifer, DEAN, Clare et CHODACKI, John, 2018. Can Richer Metadata Rescue Research? In : *The Serials Librarian*. vol. 74, n° 1-4, pp. 207-211. DOI 10.1080/0361526X.2018.1428483. Disponible à l'adresse : <https://www.tandfonline.com/doi/full/10.1080/0361526X.2018.1428483>

KEMP, Jennifer et TAYLOR, Mike, 2020. State of Open Monographs Series: Crossing the Rubicon - The Case for Making Chapters Visible. In : Digital Science [en ligne]. [Consulté le 2 mai 2020]. Disponible à l'adresse : <https://www.digital-science.com/blog/news/state-of-open-monographs-series-making-chapters-visible/>.

KOSTER, Lukas, 2020. Persistent identifiers for heritage objects. In : *The Code4Lib Journal* [en ligne]. [Consulté le 9 avril 2020]. Disponible à l'adresse : <https://journal.code4lib.org/articles/14978>.

LANGLAIS, Pierre-Carl, 2018. Les citations ouvertes, In : *Analyse I/IST* n° 28 [en ligne]. [Consulté le 18 décembre 2020]. Disponible à l'adresse : [https://www.eprist.fr/wp-content/uploads/2018/09/I\\_IST\\_28-CitationsOuvertes.pdf](https://www.eprist.fr/wp-content/uploads/2018/09/I_IST_28-CitationsOuvertes.pdf).

LAPÔTRE, Raphaëlle, 2017. Library Metadata on the web: the example of data.bnf.fr. In : *JLIS.it*. vol. 8, n° 3, p. 58-70. DOI 10.4403/jlis.it-12402. Disponible à l'adresse : <https://www.jlis.it/article/view/12402/11285>

LE PROVOST, Aline, 2020. La curation, un enjeu pour la gestion des données numériques. In : *Arabesques* n° 97 [en ligne]. [Consulté le 6 avril 2020]. Disponible à l'adresse : <https://publications-prairial.fr/arabesques/index.php?id=1793>.

LE PROVOST, Aline et NICOLAS, Yann, 2020. IdRef, Paprika and Qualinka. A toolbox for authority data quality and interoperability [en ligne]. [Consulté le 28 août 2020]. Disponible à l'adresse : <https://hal.archives-ouvertes.fr/hal-02563630>.

LINKED OPEN VOCABULARIES (LOV). *Linked Open Vocabularies (LOV)* [en ligne]. [Consulté le 16 février 2021]. Disponible à l'adresse : <https://lov.linkeddata.es/dataset/lov/>.

LO, Kyle et al., 2020. S2ORC: The Semantic Scholar Open Research Corpus, in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (ACL 2020, Online: Association for Computational Linguistics, 2020), p. 4969-83 [en ligne]. [Consulté le 28 février 2021]. DOI : <https://doi.org/10.18653/v1/2020.acl-main.447>. Disponible à l'adresse <https://www.semanticscholar.org/paper/S2ORC%3A-The-Semantic-Scholar-Open-Research-Corpus-Lo-Wang/aaad9348ffb586990bc65dd635a63826661052e0>

LUPOVICI, Catherine, 1997. L'information secondaire du document primaire : Format MARC ou SGML. In : *Bulletin d'informations de l'ABF*, n° 174. [Consulté le 16 février 2021]. Disponible à l'adresse : <https://www.enssib.fr/bibliotheque-numerique/notices/45354-l-information-secondaire-du-document-primaire>.

MARANDIN, Clarisse, 1994. « Des banques de données pour les étudiants, les enseignants et les chercheurs », Ministère de l'Enseignement supérieur et de la Recherche, 6e édition augmentée.

MARMOL, Bruno et KUNTZIGER, Bénédicte, 2019. AurÉHAL et son IdHAL rassembleur. In : *Arabesques* n° 85 [en ligne]. [Consulté le 27 mai 2020]. Disponible à l'adresse : <https://publications-prairial.fr/arabesques/index.php?id=229>.

MARTÍNEZ-GONZÁLEZ, Mercedes, ALVITE-DÍEZ, Maria-Luisa, 2019. Thesauri and Semantic Web: Discussion of the evolution of thesauri toward their integration with the Semantic Web. In : *IEEE Access*. 10.1109/ACCESS.2019.294802. [en ligne]. [Consulté le 17 avril 2020]. Disponible à l'adresse : <https://ieeexplore.ieee.org/document/8873649>.

MEADOWS, Alice, 2017. Much Ado About Metadata 2020! In : *The Scholarly Kitchen* [en ligne]. 6 septembre 2017. [Consulté le 5 mai 2020]. Disponible à l'adresse : <https://scholarlykitchen.sspnet.org/2017/09/06/much-ado-metadata-2020/>.

MEADOWS, Alice, 2019. Better Metadata Could Help Save The World! In : *The Scholarly Kitchen* [en ligne]. [Consulté le 3 mai 2020]. Disponible à l'adresse : <https://scholarlykitchen.sspnet.org/2019/06/11/better-metadata-could-help-save-the-world/>.

MEADOWS, A., HAAK, L.L. et BROWN, J., 2019. Persistent identifiers: The building blocks of the research information infrastructure. In : *Insights: the UKSG Journal*, vol. 32. DOI 10.1629/uksg.457. Disponible à l'adresse : <https://insights.uksg.org/articles/10.1629/uksg.457/>.

METADATA 2020. *Metadata 2020* [en ligne]. [Consulté le 4 mai 2020]. Disponible à l'adresse : <http://www.metadata2020.org/>.

MISSION BOTHOREL, 2020. *Pour une politique publique de la donnée* [en ligne]. [Consulté le 30 décembre 2020]. Disponible à l'adresse : [https://www.gouvernement.fr/sites/default/files/contenu/piece-jointe/2020/12/rapport\\_-\\_pour\\_une\\_politique\\_publicque\\_de\\_la\\_donnee\\_-\\_23.12.2020\\_\\_0.pdf](https://www.gouvernement.fr/sites/default/files/contenu/piece-jointe/2020/12/rapport_-_pour_une_politique_publicque_de_la_donnee_-_23.12.2020__0.pdf).

MISTRAL, François, 2016. Mutualiser les métadonnées d'autorité. L'exemple d'IdRef (Sudoc) comme projet de mutualisation de référentiels. In : POUCHOL, Jérôme (éd.), *Mutualiser les pratiques documentaires : bibliothèques en réseau* [en ligne], 2016, Villeurbanne : Presses de l'enssib. La Boîte à outils. [Consulté le 30 octobre 2020]. ISBN 978-2-37546-092-4. Disponible à l'adresse : <http://books.openedition.org/pressesenssib/6063>.

MISTRAL, François et NICOLAS, Yann, 2017. IdRef, les autorités en conquête et en partage. In : *Arabesques* 85 [en ligne]. [Consulté le 7 avril 2020]. Disponible à l'adresse : <https://publications-prairial.fr/arabesques/index.php?id=213#bodyftn1>.

MOULAISON, Heather Lea, DYKAS, Felicity et GALLANT, Kristen, 2015. OpenDOAR Repositories and Metadata Practices. In : *D-Lib Magazine* Vol. 21,

n° ¾, [en ligne]. [Consulté le 21 mai 2020]. DOI 10.1045/march2015-moulaison.  
 Disponible à l'adresse :  
<http://www.dlib.org/dlib/march15/moulaison/03moulaison.html>.

OCHANDIANO, Jean-Luc de, DUGUÉ, Alexandra, LE COUÉDIC, Laëticia et BIZOS, Isabelle, 2020. *État des lieux et recommandations pour le soutien éditorial aux revues scientifiques du site Lyon-Saint-Étienne. Rapport détaillé - avril 2020* [en ligne]. Université Jean-Moulin Lyon 3 ; Université Lumière Lyon2 ; MSH Lyon - Saint-Étienne. [Consulté le 10 juin 2020]. Disponible à l'adresse :  
<https://hal-univ-lyon3.archives-ouvertes.fr/hal-02642651>.

OPENAIRE, 2019. The Openaire Research Graph. In : *Openaire* [en ligne]. [Consulté le 18 novembre 2020]. Disponible à l'adresse :  
<https://www.openaire.eu/blogs/the-openaire-research-graph>.

OUROUK (cabinet), 2018. Rapport de l'étude n° 4 : L'articulation des archives des établissements et de l'archive nationale pluridisciplinaire HAL. In : *ADBU – Site institutionnel* [en ligne]. [Consulté le 24 février 2021]. Disponible à l'adresse : <https://adbu.fr/competplug/uploads/2018/12/Etude-COPIST-4.pdf>.

PERONI, Silvio et SHOTTON, David, 2020. OpenCitations, an infrastructure organization for open scholarship. In : *Quantitative Science Studies*, vol. n° 1, p. 428-444. DOI 10.1162/qss\_a\_00023.

PERONI, Silvio, SHOTTON, David et VITALI, Fabio, 2017. OpenCitations. [en ligne]. 16th International Semantic Web Conference. [Consulté le 6 avril 2020]. Disponible à l'adresse : <https://essepuntato.it/papers/oc-iswc2017.html>.

PERSÉE. Qu'est-ce qu'un triplestore ? In : *Data Persée* [en ligne]. [Consulté le 29 avril 2020]. Disponible à l'adresse :  
<http://data.persee.fr/ressources/quest-ce-quun-triplestore/>.

PETTERS, Jonathan, SIEMAN, Barbara, SOKOLOVA, Dina V., STOCKHAUSE, Martina et WESTBROOK, John, 2020. The TRUST Principles for digital repositories. In : *Scientific Data*. 14 mai 2020. Vol. 7, n° 1, pp. 144. DOI 10.1038/s41597-020-0486-7.

PIERROT, Raluca, 2020. Entre intelligence artificielle et science ouverte : quelles évolutions du métier de bibliothécaire ? Retour sur le congrès de l'ADBU, 17-19 septembre 2019. In : *Arabesques* n° 96 [en ligne]. [Consulté le 10 décembre 2020]. Disponible à l'adresse : <https://publications-prairial.fr/arabesques/index.php?id=1487>.

PIQUEMAL, Laurent, 2020. L'évolution de la fonction de Coordinateur Sudoc : une opportunité pour la construction d'une politique globale des métadonnées. In : *Arabesques* n° 97 [en ligne]. [Consulté le 11 juin 2020]. Disponible à l'adresse : <https://publications-prairial.fr/arabesques/index.php?id=1792>.

POMERANTZ, Jeffrey, 2017. *Metadata MOOC* [vidéos en ligne]. [Consulté le 24 février 2021]. Disponible à l'adresse :

<https://www.youtube.com/playlist?list=PLkp3pG2Rd3yqfIn3l3V32fXG4nng9Tb-H>

RENAVILLE, François, 2016. Open Access and Discovery Tools: How do Primo Libraries Manage Green Open Access Collections? In : Arnum, Ken (dir.), 2016. *Exploring Discovery: The Front Door to Your Library's Licensed and Digitized Content*, ALA Editions, p. 233-256. Disponible à l'adresse : <https://arxiv.org/abs/1509.04524>

RILEY, Jenn, 2017. *Understanding metadata: what is metadata, and what is it for?* [en ligne]. National Information Standards Organization [Consulté le 13 mai 2020]. ISBN 978-1-937522-72-8. Disponible à l'adresse : <http://www.niso.org/publications/understanding-metadata-riley>.

SMITH-YOSHIMURA, Karen, 2020. Transitioning to the Next Generation of Metadata [en ligne], OCLC. [Consulté le 6 octobre 2020]. DOI 10.25333/RQGD-B343. Disponible à l'adresse : <https://www.oclc.org/research/publications/2020/oclcresearch-transitioning-next-generation-metadata.html>.

STUART, David, 2016a. *Practical ontologies for information professionals*. London : Facet Publishing. ISBN 978-1-78330-062-4.

TAY, Aaron, 2020. Why openly available abstracts are important - Overview of the current state of affairs. In : *Medium* [en ligne]. [Consulté le 29 septembre 2020]. Disponible à l'adresse : <https://medium.com/a-academic-librarians-thoughts-on-open-access/why-openly-available-abstracts-are-important-overview-of-the-current-state-of-affairs-bb7bde1ed751>.

TAY, Aaron, 2020. The next generation discovery citation indexes - A review of the landscape in 2020 (I). In : *Medium* [en ligne]. [Consulté le 11 novembre 2020]. Disponible à l'adresse : <https://medium.com/a-academic-librarians-thoughts-on-open-access/the-next-generation-discovery-citation-indexes-a-review-of-the-landscape-a-2020-i-afc7b23ceb32>.

THÉBAULT, Vincent, 2020. BiblioLabs, un outil au service du pilotage de l'université Paris-Saclay. In : *Arabesques* n° 96 [en ligne]. [Consulté le 10 décembre 2020]. Disponible à l'adresse : <https://publications-prairial.fr/arabesques/index.php?id=1478#bodyftn6>.

WALTMAN, Ludo, 2019. Open Metadata of Scholarly Publications - Open Science Monitor Case Study, Commission européenne [en ligne]. [Consulté le 26 février 2021]. Disponible à l'adresse : <https://op.europa.eu/en/publication-detail/-/publication/a487c2bf-fe50-11e9-8c1f-01aa75ed71a1>

WILKINSON, Mark D., DUMONTIER, Michel, AALBERSBERG, IJsbrand Jan, APPLETON, Gabrielle, AXTON, Myles, BAAK, Arie, BLOMBERG, Niklas, BOITEN, Jan-Willem, DA SILVA SANTOS, Luiz Bonino, BOURNE, Philip E., BOUWMAN, Jildau, BROOKES, Anthony J., CLARK, Tim, CROSAS, Mercè, DILLO, Ingrid, DUMON, Olivier, EDMUNDS, Scott, EVELO, Chris T., FINKERS, Richard, GONZALEZ-BELTRAN, Alejandra, GRAY, Alasdair J. G., GROTH, Paul, GOBLE, Carole, GRETHE, Jeffrey S., HERINGA, Jaap, 'T HOEN,

Peter A. C., HOOFT, Rob, KUHN, Tobias, KOK, Ruben, KOK, Joost, LUSHER, Scott J., MARTONE, Maryann E., MONS, Albert, PACKER, Abel L., PERSSON, Bengt, ROCCA-SERRA, Philippe, ROOS, Marco, VAN SCHAİK, Rene, SANSONE, Susanna-Assunta, SCHULTES, Erik, SENGSTAG, Thierry, SLATER, Ted, STRAWN, George, SWERTZ, Morris A., THOMPSON, Mark, VAN DER LEI, Johan, VAN MULLIGEN, Erik, VELTEROP, Jan, WAAGMEESTER, Andra, WITTENBURG, Peter, WOLSTENCROFT, Katherine, ZHAO, Jun et MONS, Barend, 2016. The FAIR Guiding Principles for scientific data management and stewardship. In : *Scientific Data*, vol. 3, n° 1, DOI 10.1038/sdata.2016.18. Disponible à l'adresse : <https://www.nature.com/articles/sdata201618>

ZHU, Julie, 2017. Should publishers work with library discovery technologies and what can they do? In : *Learned Publishing*, vol. 30, n° 1, p. 71-80. DOI 10.1002/leap.1079. Disponible à l'adresse : <https://onlinelibrary.wiley.com/doi/full/10.1002/leap.1079>



## TABLE DES ILLUSTRATIONS

---

Tableau récapitulatif produit par Niso des divers types de métadonnées, de leurs propriétés et usages .....	17
Schéma général des flux des métadonnées des productions scientifiques en format numérique .....	18
Notice Dublin Core d'un article dans HAL .....	24
Les principaux identifiants pour les documents, les affiliations, les auteurs, les « choses » et toutes entités.....	53
Protocole de décision de Qualinka pour établir la coréférence entre une autorité et le point d'accès d'une notice.....	76
Schéma du projet Conditor dans l'écosystème de la production scientifique française.....	77
Schéma synthétisant le workflow de Conditor.....	78
Synthèse des flux de données dans et depuis LilloA .....	84
Notice du laboratoire « Structures, propriétés et modélisation des solides » dans le référentiel de Paris-Saclay .....	85
Protocole de constitution du référentiel Chercheurs de Saclay .....	86
Échanges de données entre Bibliolabs, BiblioHAL, HAL et Orcid.....	87
Propriétés des protocoles de récupération de données depuis Crossref .....	107

# TABLE DES MATIÈRES

---

<b>SIGLES ET ABRÉVIATIONS .....</b>	<b>8</b>
<b>INTRODUCTION.....</b>	<b>13</b>
<b>PARTIE A : LES MÉTADONNÉES DES RESSOURCES SCIENTIFIQUES ET LEUR ÉCOSYSTÈME .....</b>	<b>16</b>
<b>Chapitre 1 : un écosystème complexe .....</b>	<b>16</b>
<i>Des métadonnées diverses.....</i>	<i>16</i>
<i>Des acteurs et enjeux multiples dans un flux de métadonnées.....</i>	<i>17</i>
<i>Focus : les métadonnées de ressources électroniques .....</i>	<i>20</i>
<i>Les principes Fair : quatre piliers pour une bonne gestion des métadonnées .....</i>	<i>21</i>
<i>Un arsenal d'outils pour décrire les ressources scientifiques de façon standard.....</i>	<i>23</i>
<i>Bilan d'étape .....</i>	<i>27</i>
<b>Chapitre 2 : gouvernance de la science ouverte et des métadonnées de la recherche.....</b>	<b>28</b>
<i>En France : un cadre incitatif à l'ouverture des données et un écosystème d'acteurs.....</i>	<i>28</i>
<i>En Europe, une série d'initiatives articulées pour la science ouverte ..</i>	<i>33</i>
<i>Dans le monde, un réseau d'institutions œuvrant à la structuration et à l'ouverture des métadonnées .....</i>	<i>34</i>
<i>Un écosystème international dynamique, dans lequel la France est encore peu impliquée.....</i>	<i>40</i>
<i>Vers une base de données ouverte et mondiale des publications scientifiques ? .....</i>	<i>43</i>
<i>Bilan d'étape .....</i>	<i>49</i>
<b>PARTIE B : LA GESTION DES MÉTADONNÉES DANS LES ÉTABLISSEMENTS DE L'ESR .....</b>	<b>51</b>
<b>Chapitre 3 : l'enjeu des identifiants et référentiels.....</b>	<b>51</b>
<i>Identifiants pérennes et interopérabilité sur le web sémantique .....</i>	<i>51</i>
<i>Quels critères d'identifiants de qualité ? .....</i>	<i>53</i>
<i>Identifiants et référentiels .....</i>	<i>54</i>
<i>Quelles actions sont envisagées en France pour développer l'usage des identifiants et référentiels ? .....</i>	<i>55</i>
<i>Bilan d'étape .....</i>	<i>61</i>
<b>Chapitre 4 : métadonnées et « inside-out collection » .....</b>	<b>62</b>
<i>Quelle gestion des métadonnées en archives ouvertes ? .....</i>	<i>62</i>

<i>Quelles métadonnées pour les publications scientifiques en accès ouvert ?</i> .....	68
<i>Bilan d'étape</i> .....	71
<b>Chapitre 5 : les traitements automatisés de données au service de la qualité</b> .....	<b>72</b>
<i>La curation des métadonnées des éditeurs</i> .....	72
<i>Le travail sur les données de l'ESR français et des établissements de l'ESR</i> .....	76
<i>Bilan</i> .....	88
<b>CONCLUSION</b> .....	<b>89</b>
<b>ANNEXE</b> .....	<b>93</b>
<b>BIBLIOGRAPHIE</b> .....	<b>108</b>
<b>TABLE DES ILLUSTRATIONS</b> .....	<b>121</b>
<b>TABLE DES MATIÈRES</b> .....	<b>122</b>