Le défi de l'interopérabilité entre plates-formes pour la construction de savoirs augmentés en sciences humaines et sociales.

Camille Prime-Claverie, Annaïg Mahé

Le défi de l'interopérabilité entre plateformes pour la construction de savoirs augmentés en sciences humaines et sociales

Camille Prime-Claverie - MCF - Université Paris Nanterre – Dispositifs d'Information et de Communication à l'Ere Numérique – Paris, Ile-de-France (DICEN-IDF EA4420), camille.claverie@u-paris10.fr

Annaïg Mahé - MCF - Urfist de Paris / Ecole nationale des chartes - Dispositifs d'Information et de Communication à l'Ere Numérique – Paris Ile de France (DICEN-IDF EA4420), annaig.mahe@enc.sorbonne.fr

1. Introduction

A l'ère numérique, le secteur de la recherche engendre une prolifération de contenus informatisés, à la fois des documents « traditionnels » de diffusion des résultats de recherche comme des livres, des articles, des rapports ou des actes de congrès, mais également des données brutes ou des corpus numérisés, servant de matière première pour la recherche. Dans cet univers numérique, qui facilite les échanges et le partage de l'information, garantir un meilleur accès aux résultats de recherche est un objectif qui pourrait paraître aisément réalisable.

Pourtant, depuis une décennie, le secteur de la communication scientifique traverse des mutations profondes qui se traduisent par des difficultés pour l'ensemble des acteurs à se positionner dans ce nouveau contexte : « Le partage de l'IST s'accélère dans des conditions mal définies et peu régulées » [Cnrs14]. L'information se retrouve disséminée dans différentes plateformes nées sous l'impulsion de différents types d'acteurs qui affichent des positions et intérêts parfois divergents.

Dans cet environnement largement distribué, la réalisation de l'interopérabilité devient un enjeu majeur pour un meilleur accès à l'IST permettant en outre la circulation des données entre plateformes et leur enrichissement. Cette contribution propose d'aborder la question de la circulation et du partage de la littérature scientifique en sciences humaines et sociales en France à partir de données moissonnables par le protocole OAI-PMH.

Nous nous intéressons à la mise en place de l'interopérabilité dans ce domaine. Etant donné la variété des communautés professionnelles et disciplinaires mettant à disposition des métadonnées en sciences humaines et sociales, nous cherchons à évaluer l'impact de leur positionnement sur la forme et la nature des métadonnées accessibles et au final sur les niveaux d'interopérabilité et l'enrichissement sémantique ainsi rendus possibles. Nous essayons de mettre

en exergue ce qui constitue des opportunités ou des freins pour la réutilisation, l'éditorialisation et la construction de savoirs augmentées dans ce domaine.

Notre terrain se centre sur l'étude de cinq plateformes françaises mettant à disposition des documents scientifiques dans le domaine des SHS et sur l'étude d'un fournisseur de services proposant des fonctionnalités d'enrichissement. Nous nous interrogeons sur les différences entre plateformes : protocoles et standards utilisés ; niveaux d'interopérabilité organisationnelle, technique et sémantique entre ces plateformes. Enfin, nous présentons les enjeux et les limites pour l'intégration/interaction dans un autre dispositif.

2. Modèles d'interopérabilité pour la circulation des métadonnées documentaires

La question de circulation et de l'échange des métadonnées documentaires a été envisagée par différentes communautés professionnelles, ce qui explique la coexistence de différents modèles d'interopérabilité que l'on oppose parfois.

Le premier modèle d'interopérabilité documentaire s'est développé dès les années soixante dans l'univers des grandes bibliothèques. Il s'agissait à l'époque de mutualiser les efforts de catalogage et donc de permettre l'échange de notices documentaires entre catalogues de bibliothèques. Les discussions menées au sein de l'IFLA (Fédération internationale des associations de bibliothécaires et d'institutions) ont porté dans un premier temps sur la normalisation de la présentation des notices de catalogage et ont abouti sur la publication de la première norme internationale de description bibliographique (ISBD) en 1971. Plusieurs versions successives ont suivi. La dernière version dite « consolidée » car elle fédère les recommandations pour différents médias (monographies, publications en série, musique, ressources électroniques, etc.) date de 2011 et reste aujourd'hui en vigueur. Parallèlement, dès les années soixante-dix, des travaux ont porté sur la représentation informatique des informations de catalogage, donnant naissance aux célèbres formats MARC. Puis dans les années quatre-vingt, grâce au développement des réseaux informatiques, l'échange informatisé des notices a été envisagé. Né sous l'impulsion de la bibliothèque du Congrès au Etats-Unis, le protocole informatique Z39-50 devenu norme ISO en 1997 a permis la transmission des notices sur les réseaux. Il s'agit d'un protocole d'interrogation de bases de données bibliographiques développé suivant un mode client-serveur qui permet uniquement la recherche et la consultation d'informations. L'insertion, la modification ou la suppression des données ne sont pas prévues par ce modèle. Ainsi tout utilisateur équipé d'un logiciel client Z39.50 peut interroger simultanément un ou plusieurs catalogues bibliographiques accessibles par un serveur Z39.50. Ce protocole permet la formation de requêtes complexes pouvant combiner plusieurs champs en fonction des possibilités offertes par les serveurs. Les notices ainsi sélectionnées sont renvoyées au logiciel client dans un standard MARC. Cette technologie implémentée dans de nombreux systèmes intégrés de gestion de bibliothèque (SIGB) permet l'intégration rapide de notices de documents déjà catalogués par d'autres bibliothèques.

Un second modèle d'interopérabilité documentaire, plus récent, a été développé dans le contexte du mouvement prônant le libre accès aux savoirs scientifiques (Open Access). A la suite des premières initiatives de mise à disposition des textes scientifiques notamment par la création des premières archives ouvertes, s'est posé le problème de la mise en visibilité des ressources librement accessibles. Dans cet environnement, le nouveau modèle d'interopérabilité proposé prévoit des échanges de métadonnées entre deux familles d'acteurs :

– des fournisseurs de données ou encore appelés entrepôts, véritables réservoirs de documents comme les archives ouvertes par exemple ;

– et des fournisseurs de service qui moissonnent les métadonnées des premiers et proposent des applications à valeur ajoutée (recherche fédérée par exemple)

C'est dans ce cadre, que le protocole OAI-PMH (« Open Archives Initiative Protocol for Metadata Harvesting ») [VaLa00] a été initié dès la fin des années 1990 par l'Open Archives Initiative, une communauté scientifique militant à l'origine pour la prépublication en libre accès des articles scientifiques sur le web.

Comme le protocole Z39-50, ce protocole informatique prévoit la récupération de métadonnées. Il répond toutefois à une ambition très différente. Il ne s'agit pas cette fois, de récupérer quelques notices en particulier pour, par exemple, s'affranchir du travail de catalogage des nouveaux documents acquis par un service documentaire ; mais plutôt de collecter l'ensemble des métadonnées des documents d'un entrepôt ou éventuellement des derniers dépôts afin d'alimenter un service qui mettra en valeur cette collection. Ainsi, l'interrogation par le protocole OAI-PMH ne prévoit pas la possibilité de sélectionner des notices en fonction de critères précis. La collecte des informations s'effectue soit pour toute la collection, soit une période donnée et/ou pour un sous-ensemble (set) proposé par l'entrepôt. Le protocole qui s'appuie sur le protocole http et sur des standards couramment utilisés sur le web (xml, url, Dublin Core[1]) définit la manière de construire les requêtes au moyen de six verbes. Les données à collecter sont renvoyées sur la forme d'un flux xml. Un format est imposé pour l'encodage des notices récupérées : le Dublin Core, toutefois chaque entrepôt est libre de présenter ses notices dans d'autres formats. A la différence du protocole Z39-50, le protocole OAI-PMH n'interroge pas directement la base de données de l'organisme, mais un entrepôt obtenu par reformatage des données de la base. L'entrepôt OAI est donc une extension de la base d'origine, une image à un moment donnée de la base. Les données qui s'y trouvent ne sont exactement celles consignées dans la base d'origine. Il existe un décalage plus ou moins important en fonction de la fréquence de mise à jour de l'entrepôt OAI. On parle alors de recherche d'information asynchrone.

Comme ce protocole n'impose aucune exigence sur la façon dont les métadonnées doivent être stockées et organisées au cœur des systèmes [WiBN10] et prévoit des exigences minimales

[1] http://dublincore.org/

pour les formats d'export, il a connu un franc succès. Son implémentation a largement dépassé la communauté du libre accès, des archives ouvertes et des sites de dépôt de documents scientifiques, de nombreux organismes producteurs de métadonnées l'ayant choisi depuis : bibliothèques, plateformes d'éditeurs, organismes culturels, etc. L'OpenDOAR, un annuaire des entrepôts OAI-PMH, en recense actuellement plus de 2600. Le Registry of Open Access Repositories (ROAR) en recense lui plus de 3800 et, si la grande majorité des dépôts concernent des documents scientifiques, la diversité des types d'institutions concernées, des types de documents/données ainsi entreposés, ou des entrepôts eux-mêmes (notamment en volume de contenus ou en plateformes logicielles) est considérable.

Un troisième modèle d'interopérabilité né dans le contexte du web de données propose de modéliser les objets documentaires au sein d'ontologies et de relier les ressources entre elles au moyen de liens typés (Linked Data) [BeIP13, HaKl10, Nils10]. Ce modèle s'appuie sur différents standards informatiques du web sémantique (XML, RDF, OWL). L'enjeu étant de combiner et de faire interférer les données disponibles, de formuler des requêtes qui ne seraient pas prévues par les systèmes documentaires traditionnels et d'en déduire de nouvelles informations.

Plusieurs études portant sur l'interopérabilité dans les bibliothèques numériques ont montré que le modèle le plus utilisé est le protocole OAI-PMH, et plus particulièrement dans les bibliothèques académiques, loin devant d'autres protocoles tels que Z39-50, devenant ainsi la référence pour l'interopérabilité des sites de dépôts [Lopa10, Ma07, Shea14]. Ces études recensent également une grande variété parmi les formats de métadonnées : beaucoup de schémas locaux sont ainsi développés en réponse à des besoins spécifiques mais quelques schémas majeurs sont également très répandus, notamment le format MARC ainsi que le format Dublin Core (non qualifié ou qualifié) [Lopa10, Ma07, PaTo10]. Cette hétérogénéité de formats mais également dans l'implémentation d'un même format peut amener à penser que la solution résiderait dans l'adoption d'un modèle unique [AlSR12] et « de repenser à la source les modes de production de ces dispositifs selon des procédures partagées (…) » [BeCh10].

3. Présentation du terrain

Notre choix s'est porté sur l'étude des principales sources offrant des documents en texte intégral dans le domaine des SHS en France. En effet, depuis l'arrêt de la mise à jour de la banque de données Francis produite par l'INIST au 31 décembre 2014, la question de l'accès à la littérature scientifique en sciences humaines et sociales devient un enjeu majeur. Notre terrain est composé de cinq plateformes françaises mettant à disposition des documents scientifiques dans le domaine des SHS brièvement décrites ci-dessous et d'un fournisseur de services, Isidore[2],

[2] http://www.rechercheisidore.fr/

développé par HUMA-NUM, qui a pour projet de proposer un « accès unifié aux données et services numériques de SHS ».

– Cairn.info[3]

Créée en 2005 à l'initiative de quatre maisons d'édition (La Découverte, Belin, De Boeck, Erès) et soutenue par différentes institutions publiques, dont la Bibliothèque Nationale de France, Cairn.info est une plateforme permettant la diffusion sur abonnement du texte intégral de plus de 450 revues, d'ouvrages et d'encyclopédies spécialisées dans le domaine des sciences humaines et sociales.

– HAL-SHS[4]

Développé par le CCSD (Centre pour la communication Scientifique Directe) et lancée en 2005, HAL-SHS est le portail spécialisé en sciences humaines et sociales de l'archive ouverte pluridisciplinaire HAL. Comme il est indiqué sur le site, il « est destiné au dépôt et à la diffusion d'articles scientifiques de niveau recherche, publiés ou non, et de thèses, émanant des établissements d'enseignement et de recherche français ou étrangers, dans toutes les disciplines des sciences humaines et de la société ».

– Persée[5]

Créé en 2005 par le ministère de l'Éducation nationale, de l'Enseignement supérieur et de la Recherche, Persée est un portail web permettant la lecture d'articles de revues anciennes spécialisées en sciences humaines et sociales par la numérisation de leur texte intégral.

– Revues.org[6]

Créé en 1999, le portail Revues.org est l'un des services offerts par le Cléo (Centre pour l'édition électronique ouverte) soutenu par le CNRS, dont la mission est d'accompagner les éditeurs dans le développement, la structuration et l'hébergement d'une version numérique de leurs revues. Ce portail propose l'accès, la recherche et la lecture du texte intégral des revues participantes.

– Spire[7]

Spire est l'archive ouverte institutionnelle de Sciences Po.

[3] http://www.cairn.info/

[4] https://hal.archives-ouvertes.fr

[5] http://www.persee.fr/

[6] http://www.revues.org/

[7] https://spire.sciencespo.fr/

Sur un plan méthodologique, notre étude se centre sur l'analyse des informations proposées et de leur structuration, par l'interrogation directe des entrepôts OAI et par la consultation des documentations techniques mises à disposition par les plateformes. Nous nous intéressons à la prise en compte des différents niveaux d'interopérabilité par ces plateformes : interopérabilité organisationnelle, technique et sémantique.

4. Les différents niveaux d'interopérabilité

4.1 Interopérabilité Organisationnelle

Les différences organisationnelles qui apparaissent sont inhérentes aux objectifs initiaux d'élaboration des plateformes et aux modes de développement de celles-ci. Si le point commun entre ces différents portails est la mise à disposition de documents scientifiques en sciences humaines et sociales, différentes motivations coexistent : l'auto-archivage par la communauté scientifique des documents de recherche pour les plateformes HAL et SPIRE, la valorisation du patrimoine scientifique par la numérisation et la mise en ligne des collections anciennes par l'équipe de Persée, l'accompagnement à l'édition numérique et à la diffusion de documents sur des plateformes dédiées pour Revues.org et Cairn. Le tableau 1 montre la diversité des acteurs contribuant à l'alimentation de ces sources : chercheurs, professionnels de l'information, éditeurs, etc. Il en résulte des critères de formation des « sets » très différents d'une plateforme à l'autre comme le montre le tableau 2. Les portails plutôt orientés « édition électronique » comme Cairn, Revues.org ou Persée vont privilégier une offre de sets par type de document (sets d'ouvrages, de revues, etc.) et vont parfois proposer de manière plus précise la formation de sets pour chaque titre de revue, chaque titre de collection ou chaque éditeur. Seul le portail Persée propose un découpage thématique de sa collection.

Source	URL Entrepôt OAI	Type de source	Principaux contributeurs
CAIRN	http://oai.cairn.info/oai.php	Plateforme d'éditeurs commerciaux	Editeurs
HAL	https://api.archives-ouvertes.fr/oai/halshs	Archive ouverte nationale	Chercheurs, professionnels de l'information
Persée	http://oai.persee.fr/oai	Portail de revues scientifiques en libre accès (SHS)	Equipe Persée
Revues.org	http://oai.openedition.org/	Portail de livres et de revues en sciences humaines et sociales	Editeurs
SPIRE	http://spire.sciencespo.fr/dissemination/oaipmh2-no-prefix-publications.xml	Archive ouverte institutionnelle (Sciences Po)	Chercheurs, professionnels de l'information

Tab 1 : *Présentation des entrepôts OAI des cinq sources d'information étudiées*

Les archives ouvertes quant à elles présentent une construction de sets plus variés : par discipline, par type de document, et par « collection ». La notion de collection volontairement floue, permet de créer un set pour chacune des entités le demandant. Ainsi, il existe sur HAL des collections rassemblant les documents d'un laboratoire, d'un colloque, d'une conférence, d'un projet, d'un établissement, etc.

Lorsqu'ils correspondent à des objets documentaires bien déterminés, comme les sets regroupant les articles d'une revue par exemple, ceux-ci peuvent être bien décrits notamment par l'utilisation de métadonnées au format Dublin Core (cf. tableau 2).

Ces logiques de découpage, propres à chaque plateforme, rendent difficile l'extraction de sous-collections bien définies comme la constitution d'un corpus pour une discipline donnée. Elles apportent néanmoins des informations complémentaires pour les documents qui s'y rapportent.

Sources	Nombre de sets	Critères de formations des sets	Niveau de description des sets
CAIRN	899	Type de document : sur deux niveaux (ouvrages, que sais-je, repères, encyclopédie), par revue (469) ; par collection d'ouvrages (377)	Les métadonnées associées aux sets sont assez pauvres (pas d'URI et URN), description succincte du set.
HAL-SHS	4120	Type de document : 17 types différents (livre, chapitre d'ouvrage, article de journal, etc.) Discipline/Sujet : 13 Collection : 4089 collections (institution, laboratoire, conférence, projet, etc.)	Aucune description des sets. Les sets comportent uniquement les éléments nécessaires à leur déclaration : setName (nom du set), setSpec (identifiant du set).
Revues.org (Open edition)	1904	Type de document : journaux (sets par revue (413) ; livres (set par éditeur 49) ; blogs (1448) ; évènements.	Les sets sont très bien documentés par des métadonnées au format Dublin Core. Présence d'identifiants (eissn et issn) pour les revues et pour la plupart des blogs. Métadonnées au format Dublin Core utilisées pour la description des sets : dc:identifier (URI et URN) ; dc:description (description du set sur plusieurs lignes le plus souvent en anglais et en français, parfois en espagnol), dc:language, dc:publisher, dc:subject (pour les revues et les blogs)
Persée	207	Type de document : 173 revues ; 12 fonds spécialisés ; Disciplines : 22	Utilisation de quelques éléments de métadonnées Dublin Core : Différents identifiants du set (l'issn lorsque le set correspond à une revue, l'url et le doi du set), son titre, sa couverture temporelle, une description accompagnée d'une image de la première de couverture lorsqu'il s'agit d'une collection (revue, fonds) et ses droits d'exploitation.
SPIRE	50	Discipline/Sujet : (sur deux niveaux : 3 sets en sciences de l'environnement, 27 en SHS, 13 en Informatique). Quelques sets par entité (laboratoires, formations, etc.)	Aucune description des sets (contient uniquement les balises nécessaires setName, setSpec)

Tab 2 : *Critères de formation des sets proposés*
D'après le moissonnage des sources effectué le 14 janvier 2016 (requête ListSets)

4.2 Interopérabilité Technique

Une grande variété de formats proposés

Dans le cadre de la mise en œuvre du protocole OAI-PMH, le schéma de métadonnées Dublin Core est un format requis. Ce format de métadonnées proposé par un groupe de travail pluridisciplinaire et devenu norme en 2003, prévoit un jeu de quinze éléments pour caractériser toute ressource électronique disponible sur la toile (document textuel, image, son, vidéo, etc.). Ces quinze éléments de base sont considérés comme un dénominateur commun et sont librement interprétés par les différentes communautés qui les utilisent. Aussi, afin de rendre plus précise la description par le Dublin Core, les éléments de base peuvent être « raffinés » par l'utilisation d'éléments plus ciblés dits « qualificatifs »: on peut par exemple utiliser l'élément « abstract » pour préciser l'élément « description ». Il est également possible d'indiquer le recours à un référentiel pour renseigner les valeurs d'un élément. Ces deux possibilités recouvrent ce que l'on appelle le Dublin Core qualifié.

Pour les cinq entrepôts de l'étude, le tableau 3 récapitule les listes de formats retournées par les requêtes OAI-PMH utilisant le verbe ListMedataFormats et leurs nombres de notices associées. Concernant Persée, la liste renvoyée est erronée. En effet, les standards persee_erudit et tei ne semblent pas utilisés et ne renvoient aucune notice. Par ailleurs, la documentation technique du portail[8] indique l'utilisation de trois autres formats disponibles : les standards mods et marc pour les documents et le standard mets pour les collections (une revue par exemple). Ce tableau montre une offre variée de formats s'appuyant sur les standards internationaux du domaine de l'édition et des bibliothèques [More07] : des formats orientés description bibliographique comme le dublin Core (simple ou qualifié), le marc (en xml) et sa transcription mods dans laquelle les noms des éléments apparaissent aisément interprétables par des humains ou encore le format onix utilisé par les éditeurs commerciaux ; et des formats orientés édition électronique qui permettent la structuration du texte intégral des documents comme les formats tei et erudit. On note également, l'utilisation des formats mets et didl qui autorisent au sein d'une même notice l'emploi de métadonnées pour différents niveaux de granularité des documents.

[88] http://www.persee.fr/entrepot-oai

Source	Formats affichés	Nombre de notices
CAIRN	*oai_dc*	400 645
	mets	24 433
	*onix_dc**[11]***	17 356
	erudit	369 498
HAL-SHS	*oai_dc*	307 872
	xml-tei	307 872
Persée	*oai_dc*	575 302
	persee_erudit	bad argument
	tei	bad argument
Revue.org	*oai-dc*	384 192
	qdc	384 192
	mets	10 910
	tei	Accès limité aux organisations partenaires (mise à disposition du texte intégral)
	basictei	Idem
SPIRE	*oai_dc*	16 063
	mods	16 063
	didl	16063

Tableau 3 *: Liste des formats de métadonnées annoncés*
D'après le moissonnage des sources effectué le 14 janvier 2016 (requête ListMetadataFormats)

Des niveaux de granularité différents

Les spécificités culturelles des acteurs se reflètent sur les caractéristiques des données moissonnables. Ces différences s'observent non seulement par l'offre de formats de métadonnées, mais aussi par le niveau de granularité des documents pris en compte. Ainsi les plateformes d'éditeurs (Cairn, Revues.org et Persée) proposent des notices, dans différents formats, pour les collections, les titres de revue, leurs numéros et également les articles qui les composent. Le format mets, quant à lui, est utilisé pour répertorier l'ensemble des composants d'un document, par exemple les articles édités dans un numéro de revue. Il permet de faire le lien entre les différentes parties d'un document. Il n'est présenté que pour un niveau d'ensemble, un numéro de revue ou un ouvrage par exemple, ce qui explique un nombre de notices plus faible dans le tableau 3 que pour les autres formats.

Les archives ouvertes proposent des notices pour des documents qui ont été déposés volontairement. Leur rôle n'est pas de dépouiller systématiquement les parutions et d'établir les relations entre les documents issus d'une même publication. Néanmoins, un format comme didl rend possible le référencement et la description d'un document « père ».

Des propositions d'éléments différentes

Sans entreprendre une comparaison systématique des éléments de métadonnées utilisés par les différentes plateformes, nous pouvons mentionner que la variété des formats offerts permet une description approfondie des ressources bien au-delà de la proposition initiale du Dublin Core simple. Néanmoins, on observe une fois de plus que les motivations premières des acteurs conditionnent la manière d'agencer les notices : sur le choix des éléments de métadonnées et sur la façon de les renseigner. Si les données éditoriales (sources, numéro, pages, éditeurs, etc.) apparaissent systématiquement et sans ambiguïté sur les plateformes d'éditeurs, les affiliations des auteurs ne figurent pas et la description des ressources se limite en général à la présence d'un résumé. La dimension de l'écosystème scientifique est largement prise en compte par les archives ouvertes. Les auteurs, leurs organismes de recherche et parfois les projets sur lesquels ils travaillent apparaissent au sein des notices. Les documents qui sont mis à disposition par la communauté scientifique pour la communauté scientifique sont très fréquemment accompagnés de mots-clés et de résumés en différentes langues. Grâce à la modération des plateformes, les données éditoriales apparaissent dans la plupart des cas, même si elles sont parfois erronées.

4.3. Interopérabilité sémantique

L'interopérabilité sémantique s'opère d'une part, par le balisage adapté des notices, c'est-à-dire en utilisant les balises adéquates pour indiquer la valeur d'un élément, et d'autre part, par l'utilisation de référentiels communs permettant de s'accorder sur la forme ou les termes à employer pour renseigner une valeur : un format commun pour l'écriture des dates (interopérabilité syntaxique), une liste officielle pour les noms de pays, etc. Les tableaux 4 et 5 donnent des exemples de référentiels utilisés par les portails étudiés. Ils montrent aussi la

diversité des systèmes utilisés pour l'identification des ressources numériques. On note dans le tableau 4 une très faible utilisation de référentiels pour l'indexation thématique.

Source	référentiel	Exemple de valeur
HAL-SHS	Repo COAR	<dc:type>info:eu-repo/semantics/conferenceObject</dc:type>
	Repo COAR DOI	<dc:relation>info:eu-repo/semantics/altIdentifier/doi/10.3406/polix.1998.1761</dc:relation>
Revue.org	Repo COAR	info:eu-repo/semantics/article
	Repo COAR	<dc:rights>info:eu-repo/semantics/embargoedAccess</dc:rights>
	Repo COAR	<dc:date>info:eu-repo/date/embargoEnd/2018-11-26</dc:date>
SPIRE	DDC (Dewey)	<mods:classification authority="ddc">330</mods:classification>
	Repo COAR	<dc:subject>info:eu-repo/classification/jel/D84</dc:subject>
	Repo COAR	<mods:genre>info:eu-repo/semantics/book</mods:genre> grep
	Classification JEL	<mods:classification authority="info:eu-repo/authority/jel">F43</mods:classification>

Tableau 4 : Exemples de référentiels utilisés

Source	Schéma métadonnée	référentiel	Exemples métadonnée/ valeur
CAIRN	Dublin Core	URN	<dc:identifier>http://www.cairn.info/article.php?ID_ARTICLE=TE_128_0005</dc:identifier>
HAL-SHS (utilisation faible)	Dublin Core	DOI	<dc:identifier>DOI : 10.3917/rdn.408.0277</dc:identifier>
	Xml-tei	DOI	<tei:idno type="doi">10.3917/res.190-191.0073</tei:idno>
	Dublin Core	Repo COAR DOI	<dc:relation>info:eu-repo/semantics/altIdentifier/doi/10.3406/polix.1998.1761</dc:relation>
	Dublin Core dc:identifier	ISBN	<dc:identifier>ISBN : 978-2-8109-0021-3</dc:identifier>
	dc:identifier	URN	<dc:identifier>https://tel.archives-ouvertes.fr/tel-01219660</dc:identifier>
	Xml-tei	ISSN EISSN	<tei:idno type="issn">1963-1197</tei:idno> <tei:idno type="eissn">1963-1197</tei:idno> <dc:source>ISSN: 1389-5176</dc:source>
Persée	Oai_dc	DOI	<dc:identifier scheme="DOI">doi:10.3406/mots.1985.1190</dc:identifier>
	Oai_dc	URN	<dc:identifier>http://oai.persee.fr/doc/mots_0243-6450_1985_num_10_1_1190</dc:identifier>
	set	DOI	<dc:identifier>doi:10.3406/mots
	set	ISSN	1960-6001
	set	URN	http://www.persee.fr/collection/mots
Revue.org	oai_dc/dcq/Mets/	URN/DOI	<dcterms:identifier scheme="URN">urn:doi:10.4000/mots.22073</dcterms:identifier>
	<dc:identifier> <dcterms:identifier>	URI	<dc:identifier>http://mots.revues.org/22073</dc:identifier> scheme="URI">http://mots.revues.org/22073</dcterms:identifier>
	dcq/Mets	ISSN EISSN	<dcterms:isPartOf scheme="URN">urn:issn:0243-6450</dcterms:isPartOf><dcterms:isPartOf scheme="URN">urn:eissn:1960-6001</dcterms:isPartOf>

	Set Mots. Les langages du politique		<dc:identifier>uri:http://mots.revues.org</dc:identifier> <dc:identifier>urn:issn:0243-6450</dc:identifier>
SPIRE	Oai_dc	DOI	<mods:identifier type="doi">10.3917/rfs.543.0465</mods:identifier>
	Oai_dc	URI	<dc:identifier>http://spire.sciencespo.fr/hdl:/2441/9382</dc:identifier> <dc:identifier>info:hdl/2441/9382</dc:identifier>
	Oai_dc Mods	URN/ISSN URN/EISSN	<dc:identifier>urn:ISSN:13814338</dc:identifier> <mods:identifier type="eissn">13600443</mods:identifier>

Tableau 5 : Exemples d'identifiants pérennes

4. Intégration et enrichissement des métadonnées dans Isidore

Développé dans le cadre du TGE Adonis à partir de 2009, et fusionné dans l'infrastructure de recherche Huma-Num, Isidore propose une interface unique permettant une recherche fédérée de différentes sources et ressources en sciences humaines et sociales : des publications scientifiques mais également différents objets résultant des activités de recherche comme des billets de blogs, des événements, des images, des corpus de recherches numérisés, etc. Ce service collecte des informations issues de plus de 3559 sources de données numériques disponibles sur le web au moyen de différentes technologies comme les flux RSS ou le protocole OAI-PMH. L'apport de ce service est l'enrichissement sémantique des notices retrouvées par l'affectation de mots-clés « suggérés » issus de différents vocabulaires spécifiques. Ainsi, les notices qui apparaissent sans indexation thématique par mots-clés, comme l'ensemble des ressources de Cairn par exemple, peuvent être retrouvées par navigation thématique dans différents univers sémantiques.

Pour ce faire, les données récupérées, notamment celles moissonnées par le protocole OAI-PMH au format XML sont converties au format RDF, permettant ainsi la constitution d'un graphe sémantique rendant possible l'exécution de requêtes complexes et le développement de possibilités de navigation variées.

La réalisation de ce projet dans toute sa dimension rencontre en réalité un certain nombre de freins que nous énumérons ci-dessous :

– Le choix technique de ne moissonner que des notices au format Dublin Core limite le niveau de structuration des métadonnées et les informations pouvant être récupérées. C'est particulièrement dommage, car la prise en compte de certains formats actuellement disponibles chez les fournisseurs de données pourrait considérablement enrichir le graphe sémantique en créant notamment des liens entre les documents. Le format mets, en est un bon exemple car il permettrait de relier des documents à leur revue.

– Malgré les recommandations publiées par l'équipe d'Huma-Num dans un guide de bonnes pratiques [Huma14] préconisant l'emploi d'identifiants de ressources pérennes, et comme le montre le tableau 5, l'usage d'identifiants "standardisés" en particulier dans les notices en Dublin Core n'est pas stabilisé, ce qui rend difficile la possibilité d'identifier des doublons et d'enrichir les notices en conséquence.

– En pratique, peu de référentiels sont utilisés dans les ressources pour décrire le contenu thématique des ressources. En ce sens, le travail d'enrichissement sémantique d'Isidore est intéressant. Il pourrait être cependant être davantage ciblé par un travail de sélection des référentiels pertinents en fonction de l'origine des ressources par exemple. En effet, est-il intéressant d'utiliser les descripteurs issus d'un vocabulaire spécialisé en archéologie pour décrire des ressources en sociologie ?

5. Conclusion

Comme nous le mentionnions en début de cet article, les ressources scientifiques se développent dans un environnement particulièrement distribué et le défi de l'interopérabilité doit permettre d'éviter que ces ressources soient enfermées dans des silos [FoRi08, VaNe15] mais aussi que le paysage de l'interopérabilité ne devienne pas définitivement « chaotique, confus et complexe » [SuSh15]. L'étude de l'interopérabilité documentaire en Sciences Humaines et Sociales par le protocole OAI-PMH montre une richesse des métadonnées associées aux ressources bien au-delà du cadre minimal prévu, le format Dublin Core. Directement renseignées par des professionnels de l'édition ou modérées par des professionnels de l'information, ces informations associées aux ressources apparaissent fiables. De fait, il n'apparaît ni souhaitable ni nécessaire de converger vers un format commun ou de se limiter à un plus petit dénominateur commun pour assurer l'interopérabilité, mais cela ne signifie pas non plus de ne pas répondre à la « forte demande pour la fourniture d'accès unifié à de multiples entrepôts distribués et autonomes »[HaKl10]. En effet, imposer trop de contraintes aux différents fournisseurs de données peut être contre-productif, leurs motivations premières et leur expertise ne pouvant toujours s'adapter à des exigences "idéales" de convergence. Le développement d'une infrastructure d'accès unifié aux ressources doit s'appuyer sur la richesse apportée par cette hétérogénéité plutôt qu'en cherchant à la réduire [AlSR12] et cette complexité n'est pas non plus incompatible avec une simplicité d'usage : gérée en « back office », elle devient invisible pour l'utilisateur [AlSR12, SuSh15, VaNe15]. La circulation des données structurées dans différents

formats par le protocole OAI-PMH apparait comme une formidable opportunité d'alimenter un « web de données ». Des fournisseurs de service, tels qu'ISIDORE, se sont déjà engagés dans cette démarche. Un des enjeux majeurs pour les fournisseurs de données est d'utiliser des identifiants pérennes pour les documents, les ressources (DOI, Ark, ISSN etc.) et pour les acteurs de la communauté (ORCID, IdRef) afin de contribuer à la construction du Linked Science sur lequel les fournisseurs de service pourront s'appuyer.

Bibliographie

[AlSR12] ALEMU, GETANEH ; STEVENS, BRETT ; ROSS, PENNY: Towards a conceptual framework for user-driven semantic metadata interoperability in digital libraries. In: *New Library World* Bd. 113 (2012), Nr. 1/2, S. 38–54

[BeCh10] BESTER, EMMA ; CHARTRON, GHISLAINE: Difficile convergence des archives ouvertes en SIC (2010)

[BeIP13] BERMES, EMMANUELLE ; ISAAC, ANTOINE ; POUPEAU, GAUTHIER: *Le Web sémantique en bibliothèque*. Paris : Electre : Éd. du Cercle de la Librairie, 2013 — ISBN 978-2-7654-1417-9

[Cnrs14] CNRS-DIST: *Actes du colloque « Innovation et gouvernance de l'IST dans l'ESR ». Pour une politique nationale des Usages de l'IST*. Meudon : CNRS-Direction de l'information scientifique et technique, 2014

[FoRi08] FOULONNEAU, MURIEL ; RILEY, JENN: *Metadata for digital resources: implementation, systems design and interoperability*, *Chandos information professional series*. Oxford : Chandos Pub, 2008 — ISBN 978-1-84334-301-1

[HaKl10] HASLHOFER, BERNHARD ; KLAS, WOLFGANG: A Survey of Techniques for Achieving Metadata Interoperability. In: *ACM Comput. Surv.* Bd. 42 (2010), Nr. 2, S. 7:1–7:37

[Huma14] HUMA-NUM: *Guides de bonnes pratiques. Comment contribuer à Isidore avec ses données numériques ?* URL http://www.huma-num.fr/sites/default/files/guide-isidore.pdf. - abgerufen am 2016-02-15

[Lopa10] LOPATIN, LAURIE: Metadata Practices in Academic and Non-Academic Libraries for Digital Projects: A Survey. In: *Cataloging & Classification Quarterly* Bd. 48 (2010), Nr. 8, S. 716–742

[Ma07] MA, JIN: *Metadata, SPEC Kit 298* : Association of Research Libraries, 2007

[More07] MOREL-PAIR, CATHERINE: Métadonnées et XML. Des standards efficients de l'environnement numérique. In: *Ingénierie des Systèmes d'Information* Bd. 12 (2007), Nr. 2, S. 9–39

[Nils10] NILSSON, MIKAEL: *From interoperability to harmonization in metadata standardization designing an evolvable framework for metadata harmonization*. Stockholm : Skolan för datavetenskap och kommunikation, Kungliga Tekniska högskolan, 2010 — ISBN 978-91-7415-800-7

[PaTo10] PARK, JUNG-RAN ; TOSAKA, YUJI: Metadata Creation Practices in Digital Repositories and Collections: Schemata, Selection Criteria, and Interoperability. In: *Information Technology and Libraries* Bd. 29 (2010), Nr. 3, S. 104–116

[Shea14] SHEARER, KATHLEEN: *Towards a Seamless Global Research Infrastructure Report of the Aligning Repository Networks Meeting* : COAR, 2014

[SuSh15] SUMMANN, FRIEDRICH ; SHEARER, KATHLEEN: *COAR Roadmap Future Directions for Repository Interoperability*, 2015

[VaNe15] VAN DE SOMPEL, HERBERT ; NELSON, MICHAEL L.: Reminiscing About 15 Years of Interoperability Efforts. In: *D-Lib Magazine* Bd. 21 (2015), Nr. 11/12

[WiBN10] WITTEN, I. H ; BAINBRIDGE, DAVID ; NICHOLS, DAVID M: *How to build a digital library*. Burlington, MA : Morgan Kaufmann Publishers, 2010 — ISBN 978-0-08-089039-5