Identifier dans l'écosystème informationel / Septembre 2017

Diplôme national de master

Domaine - sciences humaines et sociales

Mention – sciences de l'information et des bibliothèques

Spécialité – archives numériques

Identifier dans l'écosystème informationnel,

Une réflexion autour des approches d'identification et leurs problématiques économiques, techniques et culturelles

Laure Fabre

Sous la direction de Clément Oury
Responsable données, réseaux et standards ISSN – Centre International de l'ISSN

Remerciements

Mes remerciements vont tout d'abord à mon directeur de mémoire, Clément Oury, qui m'a suivie tout au long de cette recherche et qui a fait montre d'un engagement et d'un soutien très appréciable. Merci pour tous vos conseils et votre patience.

Un grand merci à Romain Wenz, responsable du portail France Archives de la Direction des Archives de France, qui a eu la gentillesse d'accepter un entretien téléphonique pour répondre à mes questions. Votre professionnalisme et votre amabilité m'ont vraiment mise à l'aise et vos réponses ont permis à mes recherches d'être plus complètes.

Merci également à Céline Guyon pour m'avoir aiguillé sur le début du projet de mémoire et apporté ses conseils avisés. Ce premier point de vue m'a permis de vraiment cerner dans quoi je m'engageais.

Merci aussi à Emmanuel Milcent, tuteur de mon stage à l'Agence de l'eau Loire-Bretagne, et Olivia Même, stagiaire collègue dans ce même établissement. Le travail mené ensemble m'a beaucoup enrichie, et ce mémoire n'aurait pu se faire sans l'expérience de stage que j'ai eu auprès de vous.

Je remercie grandement Anne-Marie Millet pour le temps passé à la relecture et à ses suggestions très pertinentes, ainsi que son indéfectible optimisme. Tu me pousses toujours vers le meilleur depuis ma naissance.

Enfin, merci à Maxime Vigne et Alain Fabre pour leur soutien au quotidien. Votre patience et vos encouragements sont précieux.

Résumé : Le terme « identification » comporte trois axes d'interprétation : un moyen de singulariser, un moyen de contextualiser et/ou un moyen de localiser. Les pratiques d'identification dans les écosystèmes informationnels sont diverses et parlent de manière sous-jacente des objectifs des structures qui les mettent en place : elles conditionnent l'appréhension, la compréhension et l'accès aux ressources que ces structures gèrent. Dans ce mémoire, nous proposons une exploration des pratiques d'identification en interrogeant deux types de pratiques: les pratiques physiques et les pratiques numériques. Ces deux approches permettent de mettre en lumière les aspects culturels, économiques et techniques inhérents au domaine, et participent à apporter des éléments de réponse à la problématique globale que nous posons : comment identifier au mieux ?

Descripteurs : Identification, Identifiant physique, Identifiant numérique, PID, Données liées, Modèle de données, Web de données, Web sémantique, Cote, Cotation, Données, Métadonnées, Open Access, Open Data

Abstract : « Identifiying » could be implying three different meanings : individualising, contextualizing and/or locating. There is a very broad range of identification practices in information ecosystems and it speaks a lot about how the different cultural organizations integrate, comprehend and give access to the resources they manage. In this report, we will explore identification practices through two different types of approach : physical identification and digital identification. These two groundworks will allow us to enlighten cultural, economics and technical aspects, along with some piece of answers to the global problematic that we are questionning : how can we identify at best ?

Keywords : Identification, Physical identifier, Digital Identifier, PID, Linked data, Data model, Web data, Semantic web, Classification mark, Data, Metadata, Open Access, Open Data

Droits d'auteurs

Sommaire

Sigles et abréviations

AAA – *Anyone can say Anything about Anything*
ABES – Agence Bibliographique de l'Enseignement Supérieur
AFIS – Association Française d'Ingénierie Systèmes
AFNIC – Association Française pour le nommage Internet et la Coopération
API – *Application Programming Interface*
ARK – *Archival Resource Key*
BIBFRAME – *Bibliographic Framework Initiative*
BICI – *Book Item and Component Identifier*
BnF – Bibliothèque nationale de France
CAE – Compositeur-Auteur-Editeur
CCSD – Centre pour la Communication Scientifique Directe
CDD – Classification Décimale de Dewey
CDL – *California Digital Library*
CIEPS – Centre International de l'ISSN
CNRI – *Corporation for National Research Initiatives*
CQL – *Contextual Query Language*
CSI – *Code Structure Identifier*
CSV – *Comma Separated Values*
CTI – *Component Type Identifier*
CURIE – *Compact URI*
CWA – *Closed World Assumption*
DC – Dublin Core
DCMI – *Dublin Core Metadata Initiative*
DDEX – *Digital Data Exchange*
DNS – *Domain Name System*
DOI – *Digital Object Identifier*
DPI – *Derivative Part Identifier*
DPLA – *Digital Public Library of Australia*
EAD – *Encoded Archival Description*
EAN – *European Article Numbering*
EDM – *Europeana Data Model*
EIDRA – *Entertainment Identifier Registry Association*
ELI – *European Legislation Identifier*
FOAF – *Friend of a Friend Ontology*
GBIF – *Global Biodiversity Information Facility*
GCS – *Global Context Symbol*

GHR – *Global Handle Registry*

GLN – *Global Location Number*, Code-lieu

HDL – Handle

HTML – *Hypertext Markup Language*

HTTP – *HyperText Transfer Protocol*

I3C – *Interoperable Informatics Infrastructure Consortium*

IANA – *Internet Assigned Numbers Authority*

ICANN – *Internet Corporation for Assigned Names and Numbers*

IDF – *International DOI Foundation*

IFLA – *International Federation of Library Associations and Institutions*

IFPI – *International Federation of Phonographic Industry*

ILII – *International Library Item Identifier*

IMDB – *Internet Movie Database*

INDECS – *Interoperability of Data in E-Commerce Systems*

INDECS RDD – *INDECS Rights Data Dictionnary*

INRIA – Institut National de Recherche en Informatique et en Automatique

INSEE – Institut National de la Statistique et des Etudes Economiques

IP – *Internet Protocol*

IPI – *Interested Party Information*

IPN – *International Performer Number*

IRI – *International Resource Name*

ISA – *Interoperability Solutions for European Public Administrations*

ISAN – *International Standard Audiovisual Number*

ISAN-IA – *ISAN International Agency*

ISBN – *International Standard Book Number*

ISBN-A – ISBN Actionnable

ISCI – Identifiant International Normalisé des Collections

ISDL – *International Standard Document Link*

ISIL – *International Standard Identifier for Libraries and Related Organizations*

ISLI – *International Standard Link Identifier*

ISLI-RA – *ISLI Registry Agency*

ISMN – *International Standard Music Number*

ISNI – *International Standard Name Identifier*

ISNI-AA – *ISNI Attribution Agency*

ISNI-IA – *ISNI International Agency*

ISO – *International Standard Organization*

ISRC – *International Standard Recording Code*

ISSN – *International Standard Serial Number*

ISSN-L – *ISSN Link*

ISTC – *International Standard Text Code*

ISWC – *International Standard Musical Work Code*

IT – *Information Technology*

IUCN – *International Union for the Conservation of Nature*

JSON – *Javascript Object Notation*

KEV – *Key Encoded Value*

KOS – *Knowledge Organisation System*

LCCN – *Library of Congress Control Numbers*

LED – *Linked Enterprise Data*

LLD – *Library Linked Data*

LLD XG – *Library Linked Data Incubator Group*

LOD – *Linking Open Data*

LSID – *Life Science Identifier*

LSRS – *LSID Resolution System*

MDM – *Master Data Management*

MFI – *Media Format Identifier*

N/A – *Not Applicable/Not Available/Not Accessible*

NAA – *Name Assigning Authority*

NAAN – *Name Assigning Authority Number*

NBN – *National Bibliographic Number*

NCBI – *National Center for Biotechnology Information*

NER – *Named Entity Recognition*

NID – *Namespace Identifier*

NISO – *National Information Standard Organization*

NMA – *Name Mapping Authority*

NMAH – *Name Mapping Authority Host*

NSS – *Namespace Specific String*

OAI – *Open Archive Identifier/Open Archive Initiative*

OAI-PMH – *Open Archive Initiative – Protocol for Metadata Harvesting*

OCLC – *Online Computer Library Center*

OMG – *Object Management Group*

OPAC – *Online Public Access Catalog*

ORCID – *Open Researcher and Contributor Identifer*

OWA – *Open World Assumption*

OWL – *Web Ontology Language*

PDF – *Portable Document Format*

PI – *Place Identifier*

PID – *Persistent Identifer*

PII – *Publisher Item Identifier*

PMID – *PubMed Unique Identifier*

PPN – *Pica Production Number*

PURL – *Persistent Uniform Resource Locator*

RCR – Répertoire des Centres de Ressources

RDA/ONIX – *Resource Description and Access/Online Information eXchange*

RDF – *Resource Description Framework*

RDFS - *Resource Description Framework Schema*

RDMS – *Relational Database Management Software*

REL – *Rights Expression Language*

REST – *Representational State Transfer*

RHG – Registre Handle Global

SACEM – Société des Auteurs, Compositeurs et Editeurs de Musique

SCAPR – *Societies' Council for the Collective Management of Performers' Rights*

SCPP – Société Civile des Producteurs Phonographiques

SEAM – Société des Editeurs et Auteurs de Musique

SEO – *Search Engine Optimization*

SGBD – Système de Gestion de Base de Données

SICI – *Serial Item and Contribution Identifier*

SISAC – *Serials Industry Advisory Committee*

SKOS – *Simple Knowledge Organisation System*

SOA – *Service Oriented Architecture*

SOAP – *Service Oriented Architecture Protocol*

SPARQL – *SPARQL Protocol and RDF Query Language*

SPPF – Société Civile des Producteurs de Phonogrammes et France

SQL – *Structured Query Language*

SRU/W- Abréviation comprenant les protocoles *Search/Retrieve* et *Search/Retrieve Web*

SRW – *Search/Retrieve Web*

STD-DOI – *Scientific and Technical Data-Digital Object Identifier*

STI-Group – *Scientific and Technical Information Publishers*

SUDOC – Système Universitaire de Documentation

SUISA – Société Suisse pour les Droits des Auteurs d'œuvres Musicales

SVN – *Standard Version Number*

SWEO – *Semantic Web Education and Outreach Interest Group*

SWIFT – Société pour la Télécommunication Internationale Interbanque

TIB/UB – *German National Library of Science and Technology*

TSV – *Tab Separated Values*

UBC – *Universal Bibliographic Control*

UNA – *Unique Name Assumption*

URI – *Uniform Resource Identifier*

URL – *Uniform Resource Locator*

URN – *Uniform Resource Name*

UTF-8 – *Universal Character Set Transformation Format – 8 bits*

UUID – *Universally Unique Identifiers*

VIAF – *Virtual Authority File*

W3C – *World Wide Web Consortium*

WWF – *World Wide Fund for Nature*

XML – *Extended Markup Language*

XRI – *eXtensible Resource Identifier*

INTRODUCTION

Actuellement, la relation entre l'Homme et la technique est plutôt considérée comme une dépendance. Nombre d'études se portent sur le sujet de cette nécessité qu'ont les individus de toujours se reposer sur la technologie, avoir accès au réseau, être en permanence connectés, si bien que c'est presque perçu comme la maladie de notre siècle. C'est un peu ce que dénonce Bruno Latour, philosophe et anthropologue, quand il dit que ce qui est valorisé actuellement est la capacité de chacun à s'affranchir de cette technologie pour prouver sa valeur intellectuelle[1], comme si l'on considérait que me centre de l'intelligence ne se situait que dans la tête. Il regrette la négligence de ces moyens « humbles » que sont les outils techniques, et qui constituent selon lui les « *conduits indispensables à l'exercice de la connaissance* », ayant donc une importance décisive.

Dans la pensée de Douglas Engelbart, ingénieur américain et notamment pionnier dans l'informatique, l'Homme et la machine entretiennent des rapports de coévolution[2]. Au lieu d'imaginer une mécanique insidieuse de remplacement mutuel, il imagine la notion comme une spirale de constante évolution en parallèle, selon un principe de création-amélioration. L'Homme créé la machine qui va l'aider à lui permettre de s'améliorer, et donc de l'améliorer, et ainsi de suite. Ces deux éléments totalement hétérogènes que sont l'Homme et la machine s'articulent vers l'objectif commun d'excellence sur la base essais-erreurs. C'est également le point de vue de Joseph Licklider, informaticien qui développera notamment l'idée d'un réseau global de données, en étant le premier à utiliser les ordinateurs de manière massive et à travailler à l'analyse de la relation Homme-machines. Il est très favorable à cette symbiose (qui sera notamment le titre d'un de ses articles les plus reconnus[3]), et il visualise l'ordinateur comme un outil de modélisation de toutes les disciplines.

Hans Moravec, spécialiste de robotique et d'intelligence artificielle, développe la question avec un paradoxe : *« Il est facile de faire en sorte que les ordinateurs exhibent une intelligence d'adulte sur des tests de QI, mais difficile voire impossible de leur donner les compétences d'un enfant de 1 an quand il s'agit de perception et de mobilité*. »[4] En d'autres termes, les facultés de raisonnement sont difficiles pour les humains mais faciles pour les ordinateurs, tandis que les facultés sensorimotrices et l'adaptabilité du langage sont des propriétés sur lesquelles l'humain (et plus généralement les organismes biologiques vivants) sont pour l'instant inégalés. Les deux se complètent donc bien de manière plutôt avantageuse.

[1] LATOUR, Bruno. Dans JACOB, Christian *(dir.). Lieux de savoir, Espaces et communautés*. Albin Michel, 2007. Chap. Pensée retenue, pensée distribuée. p605-615

[2] C'est-à-dire qu'ils vont évoluer parallèlement tout en s'enrichissant l'un l'autre.

[3] LICKLIDER, Joseph. Man-computer symbiosis, *IRE Transactions on Human Factors in Electronics*, Vol. HFE-1, Laboratoire d'informatique et d'intelligence artificielle MIT, 1960. Pp 4-11

[4] VAN HOOLAND, Seth. VERBORGH, Ruben. *Linked Data for Libraries, Archives and Museums: How to Clean, Link and Publish your Metadata.* Amer Library Assn Editions, 2014. 224p.

Le relationnel Homme-machine et l'organisation de la connaissance

Les bases de l'informatique graphique trouvent leurs origines dans les années 1960 et ont été notamment portées par Engelbart et Licklider qui ont développé la théorie du relationnel interactif Homme-machine. Les surfaces d'inscription qu'étaient les toutes premières tablettes sont apparues et ont participé à développer la notion d'interface qui est aujourd'hui très commune et indispensable. Tablette d'aujourd'hui et d'hier se font écho, comme un éternel retour vers ce qui nous parle le mieux : la transcription simple et géniale de la pensée dans son état premier.

Parallèlement, le développement par Licklider de l'idée de réseaux et de la mise en lien de technologies pour créer ce qui sera plus tard Internet prolonge l'idée que les machines et la technologie sont en interaction avec nous mais également entre elles. La logique de réseau tend à explorer le potentiel d'une machine par un processus collectif, processus auquel Licklider contribuera fortement jusqu'à parvenir à mettre en lien des éléments hétérogènes. Tout comme dans un cerveau, il imagine la machine selon une dynamique de bas en haut : système, sub-système et composants. De surcroit, la possibilité nouvelle de créer des courbes, dessins, cartes et arborescences est une technique innovante de projection intellectuelle qui prolonge les capacités du cerveau via ses qualités de mémoire, de connexions étendues et de calcul. Cela permet, *in fine*, une gestion très importante de la simultanéité. La machine apporte donc, comme le dit très justement Bruno Latour, un « truchement » à l'esprit pour fabriquer des trésors d'intelligence, qui ne doit pas être négligé. Là où précédemment nous avions une logique d'apprentissage humaine, se substitue une logique de convivialité : il ne s'agit plus de savoir, mais de savoir utiliser/communiquer. Le savoir global se substitue au savoir individuel. La mise en commun des connaissances et l'utilisation conjointe d'outils intellectuels permet d'une part d'accéder à une autre dimension de réflexivité, et d'autre part fait évoluer le cerveau dans l'optique d'utiliser les données qui lui sont fournies pour produire une connaissance et une intelligence foncièrement humaine.

L'organisation de la pensée a été une des préoccupations les plus importantes des siècles derniers. L'incroyable expansion de la production de données et de documents, allant de pair avec la mondialisation et les échanges internationaux, a constitué un enjeu de gestion majeur qui a conduit au développement des sciences de l'information. Paul Otlet, dans son ouvrage *Traité de la documentation* en 1934 extrait déjà des principes fondateurs : 1. Unité, 2. Universalité, 3. Expansibilité, 4. Rationalisation, normalisation, standardisation, 5. Coopération, 6. Publicité, 7. Sériation des efforts.[5] La notion d'universalité est celle qui revient le plus souvent avec celle de l'unité (unité de la documentation, unité de la connaissance).

Ce mémoire, en allant au-delà des considérations techniques et pratiques spécifiques à l'identification que nous aborderons, tentera de manière sous-jacente d'apporter des éléments de réponse sur l'état actuel de la gestion de la connaissance mondiale, par les outils et les principes utilisés dans le paradigme actuel.

[5] OTLET, Paul. *Traité de documentation*. Liège : C.L.P.C.F, 1989. Réimpression de l'édition de 1934. Préface de Robert ESTIVALS, Avant-propos d'André CANONNE. p 374

Internet et le web

Mais retraçons tout d'abord certains éléments de contexte. En ce qui concerne les technologies de gestion de la connaissance qui nous concernent, il faut ici distinguer plusieurs terminologies qui désignent des notions différentes : Internet, Web (ou World Wide Web), le web sémantique et les protocoles d'échange.

Mirna Willer et Gordon Dunsire définissent ainsi Internet :

> « L'internet est un réseau global de réseaux locaux d'ordinateurs qui communiquent entre eux en utilisant un socle commun de protocoles. Les protocoles de communication apportent un cadre afin d'échanger de manière sécurisée et cohérente des messages et de la donnée. »[6]

Les protocoles les plus en vue qui se sont notamment répandus et qui ont perduré jusqu'à devenir des socles « communs » ont été le système DNS (*Domain Name System*), qui fonctionnait avec l'identification de chaque machine par une adresse IP (Internet Protocol), et la technologie de l'hypertexte* http (*Hypertext Transfer Protocol*), développée par Tim Berners-Lee, inventeur du « web » à proprement parler. Le terme d'hypertexte en lui-même n'est cependant pas de son fait, le concept a été créé par Ted Nelson en 1965, pour l'appliquer à son système Xanadu.

Le *World Wide Web* (WWW) ou web, quant à lui, était mis en place originellement pour les chercheurs et scientifiques du CERN avec le protocole http. Il fonctionne comme un système de référencement : inter-actionnable, il s'agit « *d'embarquer une référence à un document dans un autre document* », à l'image des renvois que l'on peut déjà trouver dans les citations scientifiques. Les codes bibliothéconomiques sont utilisés comme des rebonds hypertextuels. Le principe n'est en effet pas nouveau, les systèmes d'annotation médiévales des gloses utilisaient déjà ce mode de référence.[7] C'est donc en quelque sorte un service d'indexation de citation auto-génératif.[8] Les documents présents sur le web tiennent donc plus de la ressource intégratrice que de la publication sérielle.

Plusieurs facteurs sont responsables du succès du web de Tim Berners-Lee face à d'autres solutions et d'autres protocoles :

- La transparence de sa méthode : appelée le « *view source effect* »*. Il s'agit en réalité de la capacité que peut avoir n'importe qui surfant sur le web de consulter le code source de la page, autrement dit sa mécanique interne. Beaucoup ont pu apprendre en observant et, de fait, réutiliser. Cela explique notamment la popularité d'HTML comparé à d'autres langages de développement web.[9]
- L'uni-directionnalité des références entre les documents. Originellement, l'hypertexte était développé par Ted Nelson sur un principe bidirectionnel avec Xanadu: le document référé permettait de revenir au document le référant, dans un sens comme dans un autre. Avec http, cette simplification a été à double

---

[6] WILLER, Mirna, DUNSIRE, Gordon. *Bibliographic Information Organization in the Semantic Web, 1st Edition*. Chandos Publishing, 2013. 350p.

[7] WENZ, Romain. Dans ALIX, Yves (dir.). *Bibliothèques en France 1998-2013*. Editions du cercle de la librairie, 2013. 279p. Chapitre « Des catalogues aux métadonnées : la bibliothèque vers le Web sémantique » p.160-171.

[8] WILLER, Mirna, DUNSIRE, Gordon. *Op. Cit.*

[9] BOOTH, David. *Four Uses of a URL : Name, Concept, Web Location and Document Instance*. W3C, 2003. [en ligne] Disponible sur https://www.w3.org/2002/11/dbooth-names/dbooth-names_clean.htm [consulté le 31/03/2017]

tranchant. Si elle a permis effectivement l'expansion du web en implantant la distribution des serveurs de ressources, elle génère le « *Link maintenance problem* »*, qui anime les communautés depuis la création du web et qui n'a toujours pas trouvé réponse. Il s'agit de l'incapacité d'une référence à évoluer avec la ressource liée : qu'elle soit déplacée, supprimée et le lien devient obsolète sans que son gestionnaire en soit averti. Cela créé un nombre incommensurable de liens morts non maintenus qui participent à l'instabilité et l'inconstance du web, la bien tristement universelle « frustration de l'erreur 404 ». Ted Nelson disait d'ailleurs à ce propos ironiquement : « *la réaction de la communauté de recherche sur l'hypertexte en voyant le World Wide Web a été comme de savoir qu'elle avait un enfant complètement développé, et délinquant de surcroit.* »[10]

- Le web a toujours été imaginé, de par l'analogie que l'on peut faire entre « page », « référence bibliographique », « document », tel un livre. Cependant, son déroulement n'est justement pas linéaire, il se construit par renvoi et superpositions. Olivier Ertzscheid, chercheur en Sciences de l'Information et de la Communication à Nantes et auteur du blog très suivi *Affordances*, développe d'ailleurs l'idée que le web est un « théâtre » plus qu'un livre, les pages sont des espaces pouvant être de l'ordre du dispositif, de la relation, de la surface ou de la profondeur. « *Le web est un média palimpseste. La rature, la surcharge, les transparences lui sont consubstantielles*. » Le web permet en outre la renégociation, la re-computation, il possède une « *réelle puissance disruptive* » qui en a également fait son succès.[11]

Il y a donc d'une part, un réseau global de machine constituant un socle matériel et technique permettant leur intercommunication : Internet ; et d'autre part, une couche supérieure formée par une toile (littéralement) qui permet de stocker des documents, des « objets numériques » qui se référencent entre eux et se répondent, pas toujours de manière très fiable.

Le système DNS (*Domain Name System*) gère quant à lui l'identification des différents matériels (ordinateurs, serveurs) connectés sur le web, permettant leurs échanges. Chaque ordinateur possède un nom qui peut évoluer -statique ou dynamique- mais qui constitue son identité et son « adresse » lors d'interaction avec les autres. Il lui permet notamment de faire des requêtes et de recevoir des réponses, demander une page web, un document, un service. Ainsi, chaque système connecté au réseau a déjà sa propre adresse IP (*Internet Protocol*).

Afin que celle-ci soit plus gérable, elle est remplacée parfois par un « nom de domaine »* : cela permet une souplesse vis-à-vis des machines, si on déplace le contenu publié d'un ordinateur à un autre, on peut conserver ce même nom de domaine. Cela forme les « sites web », qui constituent la toile.

---

[10] Propos de Ted Nelson recueillis par Nick Gibbins, lors de la conférence Eighth ACM International Hypertext Conference ayant eu lieu à Southampton les 6 et 7 avril 1997, et cités dans *The eighth ACM International Hypertext Conference*, Ariadne, n°9, 1997. [en ligne] Disponible sur http://www.ariadne.ac.uk/issue9/hypertext [consulté le 13/07/2017]

[11] ERTZSCHEID, Olivier. *De quoi la page Web est-elle le nom ? L'enluminure du code*. Blog Affordances, 2011. [en ligne] Disponible sur http://affordance.typepad.com/mon_weblog/2011/03/de-quoi-page-web-est-le-nom.html [consulté le 29/11/2016]

Evolution du web et de ses objectifs

Romain Wenz, dans son chapitre dédié au web et au web sémantique dans *Bibliothèques en France (1998-2013)* identifie trois phases de « modes d'accès » au web depuis 1990[12] :

- Une simple liste d'annuaire de sites, reproduisant un peu les catalogues,
- L'explosion des moteurs de recherche à grande échelle, survenant dans les années 2000,
- Et enfin l'apparition de systèmes de préconisation personnalisés et individuels.

En outre, nous distinguons trois évolutions majeures dans l'appréhension des objectifs du web, qui sont liées à ces modes d'accès plus ou moins indirectement : le web 1.0 qui correspond au web dit « de documents » classique (échange de données entre des machines et mise en ligne de contenu) ; le web 2.0, pour évoquer l'expansion des réseaux sociaux, du participatif, du collaboratif, aussi appelé web de réseau ; et enfin le web 3.0 qui nous préoccupe, autrement dit le web sémantique ou web de données que nous développerons plus avant. Cette dénomination sous forme de *versioning** informatique est originellement le titre accrocheur d'un article Tim O'Reilly, publié en 2004, sur l'avènement des réseaux sociaux sur le web et qui a progressivement été réemployé.[13]

Les métadonnées sont la pierre angulaire du développement des moteurs de recherche. En effet, le premier algorithme de Google se base sur cela : qui cite quoi. La métadonnée est une description de la donnée (une donnée sur la donnée), telle une notice bibliographique. Ajouté au mécanisme de référence, nous voyons bien l'analogie qui peut être faite entre le monde des technologies de l'information (IT) et celui du livre et de la bibliographie qui perdure encore. L'utilisation de métadonnées n'est cependant pas nouvelle non plus, elle remonte à -3000 av J.C, avec les assyriens qui apportaient des descriptions sur les cylindres d'argiles pour en dénoter le contenu.[14]

Identifier, faire exister

Dans ce mémoire nous traiterons plus particulièrement de la question des identifiants physiques transposés au monde numérique, mais posons-nous d'abord la question : qu'est-ce qu'un identifiant ? Qu'est-ce qu'identifier ?

Si la notion d'identifiant dépasse aussi largement en terme d'ancienneté celle du web et d'internet, elles sont néanmoins assez connectées. Analyser cette liaison permet de voir les enjeux derrière l'idée même de l'identification, et a fortiori l'identification numérique.

Le web se base donc sur le principe de l'hypertexte qui est l'avènement du règne de l'immédiateté et de l'intuitivité : « *l'intégralité du matériau textuel devient une potentielle clé d'accès* »[15]. Et l'hypertexte repose directement sur la question de l'identifiant. On peut identifier sans même qu'une ressource existe : il n'y a pas de

[12] WENZ, Romain. *Op. Cit.*

[13] *Ibid.*

[14] STOCKWELL, Foster. *A history of Information storage and retrieval*. Mc Farland&Co, Jefferson, 2007. 208 p.

[15] BERMÈS, Emmanuelle, ISAAC, Antoine, POUPEAU, Gautier. *Le Web sémantique en bibliothèque*. Paris : Electre-Ed. du Cercle de la Librairie, 2013.

vérification, ce qui accorde une vraie souplesse au processus (ce qui, nous l'avons vu, n'est d'ailleurs pas étranger à son succès). L'identifiant web, que l'on appelle URI (*Unique Resource Identifier)* est au cœur de la question. Le web, quant à lui, est construit sur trois notions fondatrices qui en forment son architecture : l'identifiant, la représentation et la ressource. A chaque entité existante correspond une URI, et toute entité identifiée par une URI est une ressource.

L'identifiant conditionne donc à lui-même l'existence de quelque chose sur le web et sur internet. Tout ce qui a un nom, même s'il n'existe pas physiquement, existe par ce seul fait. Nommer numériquement c'est encore plus que désigner, c'est donner vie. Les enjeux sont grands : identifier un objet, une ressource c'est donc non seulement lui donner vie mais lui attribuer une étiquette, le définir, mais également ce qui va permettre de le retrouver, le localiser, le discriminer. Cela conditionne toute son existence à proprement parler et la possibilité même de pouvoir l'exploiter.

A partir de cela, il est intéressant de savoir comment soi-même nommer, définir et localiser ses propres ressources. A l'instar de beaucoup d'entités, les ressources ont souvent plusieurs noms et cela brouille la compréhension que l'on en a. Rappelons-nous ce qui fut infligé aux humains dans l'histoire de la fameuse tour de Babel, ambitieux et insolent projet qui vaudra à l'humanité comme punition d'être incapable de se comprendre. En effet, c'est une malédiction bien plus insidieuse que l'on ne pourrait le penser : un objet, une idée, un concept qui possède plusieurs noms ne peut pas être communiqué correctement. Ici, c'est que nous cherchons à faire, communiquer une ressource, une information, partager un savoir ou un objet, le désigner et le retrouver de façon unique, pérenne et univoque.

Le lien entre monde physique et monde numérique pose de multiples questions. En effet, les deux se transposent et se superposent à bien des égards, et il est difficile de faire parfois la part des choses entre un monde physique empli d'objets « palpables », et un monde numérique qui parle de ceux-ci de manière « impalpable ». Si auparavant le monde numérique n'était pas imaginé comme autre chose qu'un pendant du monde physique, il commence à se détacher progressivement de son enveloppe réelle pour revendiquer une existence à part entière : la copie numérique, pour peu qu'elle soit gérée dans des conditions garantissant son intégrité, sa fiabilité et son authenticité, pourra désormais remplacer et conduire à sa destruction son bon vieil original papier[16].

Cette porosité des mondes bouleverse le paradigme connu jusqu'ici et la distinction entre les deux n'est plus si évidente. Aujourd'hui, alors que le monde physique s'invite dans le monde numérique et que le monde numérique colonise tous (ou presque) les aspects du monde physique, il s'agit de créer du lien, des ponts et des passerelles entre les deux afin de converser selon un même langage, et ainsi éviter l'effet « Tour de Babel ».

[16] Décret n°2016-1673 du 5 décembre 2016 relatif à la fiabilité des copies et pris pour l'application de l'article 1379 du code civil.

Web sémantique, web de données

C'est ainsi que naquit le principe du web sémantique, en 2001. Il vit le jour des mêmes mains qui avaient autrefois accouché le web, celles de Tim Berners-Lee, explicité pour la première fois dans un article du *Scientific American Magazine*[17].

L'idée principale, que nous développerons ci-après dans une partie spécifique de notre propos, découle en réalité des recherches sur l'intelligence artificielle que nous avions évoquées plus haut. L'objectif de celles-ci est de « *permettre aux machines de naviguer dans notre langage naturel, réduire l'ambiguïté, les rendre adaptables : elles peuvent ainsi raisonner, appliquer des interprétations en fonction des liens qu'on leur donne.* » [18] Il s'agit de rajouter une nouvelle « strate » au web, un réseau sémantique supplémentaire aux données, afin d'ajouter du sens à cette entité : chaque élément est défini dans son contexte. Le langage humain fonctionne d'ailleurs de la même manière. Un mot de ne se lit qu'au sein d'un groupe, on va pouvoir interpréter son sens grâce à l'articulation des notions qui constituent la phrase. C'est tout l'intérêt de l'adaptabilité humaine, propre aux organismes biologiques et qui d'ailleurs en conditionnent la survie. Le mot seul se retrouve dénué de toute substance, voire peut mener à des quiproquos, des doubles ou contre-sens, des synonymies. Il est le jouet de l'imagination de son récepteur. Les machines, quant à elles, n'ont pas d'imagination.

Concrètement, ce mécanisme et cette « strate » de sens permet une organisation de l'information lors d'une recherche : de « *sélectionner, analyser, réorganiser dans une nouvelle interface des données choisies afin de proposer une nouvelle vision, de nouvelles perspectives, points de vue, information, etc.* »[19] Le web sémantique fournit un cadre structuré qui permet d'appréhender l'existence de concepts mobilisables par la machine, d'interagir avec l'utilisateur humain, lui recommander des entités connexes, suggérer des documents, et qui plus est reconnaitre « intuitivement » les notions qui se recoupent. Autrement dit, permettre par la liaison des données, d'acquérir comme on dit « de la jugeote ». Ces notions ne sont pourtant pas nouvelles. Le domaine des contrôles bibliographiques exploite déjà le potentiel des données liées, en traitant cette masse par le classement et la description. Le W3C et Tim Berners-Lee utilisent un modèle de données abstrait dans le but premier de gérer des métadonnées interopérables.

En outre, la plupart des relations implicites du web restent non répertoriées et non exploitées[20]. Il y a là un réel potentiel économique et scientifique : les humanités numériques et les possibilités d'appliquer le « *peer-reviewing* »* au sein même du web amorce l'idée d'une auto-gérance de la référence sur le web et les réseaux par la « sagesse du peuple ». On assiste à l'émergence de services de détection et de recherche des relations sur le web (tels que Technorati) qui ont compris l'intérêt représenté par de tels systèmes. Nous développerons notamment cette idée dans la troisième partie de ce mémoire.

---

[17] BERNES-LEE, Tim, HANDLER, James, LASSILA, Ora. The semantic web. *Scientific American Magazine*. 2001.

[18] VAN HOOLAND, Seth. VERBORGH, Ruben. *Op. Cit.*

[19] BERMÈS, Emmanuelle, ISAAC, Antoine, POUPEAU, Gautier. *Op. Cit.*

[20] WRIGHT, Alex. *Exploring a 'Deep Web' that Google can't grasp*. www.nytimes.com, 2009. [en ligne] Disponible sur http://www.nytimes.com/2009/02/23/technology/internet/23search.html [consulté le 13/07/2017]

Problématique explorée

La question qui nous intéresse dans ce mémoire est donc la suivante : comment utiliser l'identification pour optimiser les bénéfices des données liées, et s'adapter aux enjeux des paradigmes actuels ? Cela dérive sur une infinité de problématiques que nous nous efforcerons ici de limiter à trois interrogations :

- Deux questions en une : Quoi identifier ? Pourquoi identifier ?
- Quels sont les enjeux et les conséquences économiques, techniques et culturels ?
- Comment identifier au mieux, et ainsi s'adapter et se placer dans un contexte concurrentiel ?

Ayant fait le tour dans cette introduction des différentes notions et concepts que nous aborderons tout au long de ce propos, la suite de la réflexion s'articulera sur trois grands points :

Tout d'abord, nous ferons une première plongée au cœur des notions avec l'exploration de l'identifiant avant le numérique, et plus spécifiquement dans les bibliothèques, dans les archives, dans les musées et dans les circuits d'édition et de publication. Qu'avait-on besoin d'identifier, comment était-ce fait ? Quels étaient les enjeux ? Nous comparerons ainsi le traitement de l'identifiant numérique et celui de l'identifiant physique : les différences de gestion qu'ils impliquent et leur répercussion lors de réutilisation d'identifiants déjà existants et déjà attribués. Nous verrons ensuite toujours dans cette première section plus en détail les principes apportés par le numérique (l'objet identifié, les notions d'URI, d'URL, d'URN, etc.), le web sémantique/web de données ainsi que la structuration de données qui nous éclairera sur les mécanismes qui régissent les processus d'identification. Nous terminerons cette première partie par un tour d'horizon des contextes d'identification : l'architecture du web, les outils utilisés, le LED (*Linked Data Enterprise*) qui apportent un nouveau point de vue et surtout un contexte au processus d'identification.

Nous poursuivrons avec une seconde partie plus technique, spécifiquement axée sur l'identifiant même, avec une exploration de l'anatomie de l'identifiant : ses composants, ses enjeux, ses usages, ses problématiques. Nous développerons à la manière d'un exposé les grands d'identifiants présents actuellement sur le marché ainsi que leur origine, leur vocation et leur syntaxe, et ce afin d'avoir une vue globale de ce qui est faisable en terme d'identification à notre instant *t*. Puis nous aborderons les questionnements liés à l'interopérabilité entre les systèmes proposés et les méthodes de gestion des identifiants, afin de mieux comprendre comment chacun de ces acteurs évoluent et coexistent.

Enfin, nous conclurons sur une troisième et dernière partie dédiée entièrement au retour d'expérience, à la prise de distance théorique vis-à-vis de la technologie. Nous irons un peu plus en profondeur grâce à l'appui de trois études de cas : la BnF (Bibliothèque nationale de France) pour le point de vue bibliothèque, la BBC (British Broadcasting Corporation) pour le secteur de l'édition et de publication de contenu, et la Direction des Archives de France pour l'exploitation des fonds patrimoniaux. Nous complèterons ce panorama avec des commentaires, des analyses et des ajouts qui concernent d'autres structures afin de faire des mises en perspective. Nous mettrons particulièrement l'accent sur la réutilisation qui est faite des identifiants physiques dans le monde numérique, intéressante pour lier les deux et replacer le contexte de chaque institution avant le web. Pour terminer nous nous attacherons à la présentation des différentes solutions et bonnes pratiques

recommandées et exploitées, afin de dégager ce qui nous semble, de manière subjective, les choix les plus appropriés pour la mise en place d'un système d'identification au sein d'une structure privée ou publique.

1. PLONGEE AU CŒUR DES NOTIONS

1.1. L'IDENTIFIANT AU SERVICE D'UNE REALITE PALPABLE

1.1.1. Dans les bibliothèques

Classification et cotation : une symbiose

La cotation* en bibliothèque débute avec les différentes classifications, les deux concepts étant intimement liés. En effet, Il semble qu'il y ait toujours eu une frontière très poreuse dans les bibliothèques entre « classer » et « identifier » qui sont, en soi, deux opérations différentes. Les notions ont cependant été construites en parallèle, les cadres de classement évoluant et impliquant la mise en place de nouvelles formes de cotation.

Le besoin de sectoriser, classer et organiser le monde n'est pas nouveau. Le langage même possède intrinsèquement une force classificatoire, qui traduit à quel point la classification est une « *sorte d'obsession humaine* ».[21] En 1684, Nicolas Clément, garde de la Bibliothèque du roi, établit 23 divisions systématiques en classant par format et par sujet les ouvrages, qui seront utilisées jusqu'au début de l'année 1997 à la Bibliothèque nationale de France (BnF). Cette classification faisait alors l'objet d'une cotation systématique par numérotation qui, par la suite et à l'issue des évènements révolutionnaires (accroissement des fonds, réorganisation), a été modifiée successivement. Les principes de cotation se sont adaptés aux classifications, en utilisant différents systèmes tels que les lettres dites « cataloguées », les cotations transitoires, la numérotation continue par millésime, ou alors ils variaient sensiblement en fonction du type de classement (thématique, documentaire, topographique…).[22]

Les classifications dites globales, censées comprendre l'entièreté des savoirs humains, ont toujours été développées en parallèle à une identification des concepts, dans le sens de l'attribution systématique d'un identifiant. Cette dernière se devait d'être à la fois rigoureuse et univoque (les chiffres se prêtaient relativement bien à ce type de besoin) mais également de pouvoir être déployée selon des niveaux de granularité permettant l'accroissement des domaines eux-mêmes. La classification décimale Dewey (CDD), établie en 1876 par Melvil Dewey, permettait de développer une granularité thématique reposant sur une hiérarchie des concepts : 10 concepts globaux au même niveau de structure logique pouvant eux-mêmes être subdivisés hiérarchiquement 10 fois, sur 5 niveaux de granularité différents. Elle donne des indications sur le thème d'un ouvrage, mais plusieurs ouvrages peuvent avoir la même cote Dewey. Elle « identifie » donc de manière très générique sans être complètement unique. La question de l'unicité, très cruciale sur le web, est ici secondaire ; ce qui en soi n'est pas un problème dès lors que l'on est dans une identification locale à des fins de rangement. Cela est dû à l'historique des bibliothèques mais également aux caractéristiques inhérentes au livre.

---

[21] VIRY, Claude-Michel. *Guide historique des classifications de savoirs ; enseignement, encyclopédies, bibliothèques*. L'Harmattan, Paris, 2013. 256 p.

[22] BnF. *La cotation à la BnF*. 2012. [en ligne] Disponible sur http://www.bnf.fr/fr/professionnels/anx_catalogage_indexation/a.cotation_bnf.html [consulté le 14/07/2017]

Le bien informationnel et ses caractéristiques

Avant d'entrer dans le sujet des bibliothèques même, il s'agit donc de s'intéresser à l'objet qu'elles manipulent : le livre, ou plus généralement le bien informationnel*. Les bibliothèques s'occupent d'objets semi-sériels : des publications plus ou moins rares qui peuvent exister en plusieurs exemplaires, ou même de manière unique (et c'est à ce moment que la séparation entre bibliothèque et archives devient plus trouble). Elles se doivent donc de gérer, plus que des objets unitaires en soi, une connaissance globale contenue dans des objets qui pourraient être retrouvés dans diverses parties du monde, en parallèle de la gestion en amont des éditeurs et diffuseurs qui en extraient d'abord une valeur marchande. Cet aspect du livre comme « produit » (identiquement répété) dont on ne pourrait identifier que le concept général (le contenu conceptuel, la forme et l'édition, comme pour d'autres types de produits) se confronte à cette dimension « informationnelle » qui apporte les notions d'asymétrie, de bien d'expérience, de bien collectif, et de reproductibilité entre autres, caractéristiques mêmes de l'économie du livre. [23]

Concernant le bien informationnel à proprement parler, Jean-Michel Salaün distingue 7 caractéristiques propres à celui-ci : la non-destruction, l'expérience, l'interprétation, le prototype, la plasticité, l'attention et la résonnance[24]. Le bien informationnel est asymétrique car la possession d'une information utile, d'un savoir, permet l'obtention d'un avantage, qu'il soit économique, situationnel, ou financier. Sa valeur n'est connue qu'une fois consommé, car c'est un bien d'expérience. La non-destruction est en outre l'un des critères fondamental que la bibliothèque par extension représente à elle seule : des centaines de savoirs qui enrichissent ceux qui les partagent sans détruire le bien qu'ils consomment.

Par ailleurs, dans son livre *Vu, Lu Su* Jean-Michel Salaün décrit trois valeurs ajoutées apportées par les bibliothèques à ces principes : la mutualisation des documents (réutilisation par plusieurs personnes), la *sérendipité* (les découvertes pouvant être faites par le rassemblement de documents sélectionnés à travers des collections), et le gain en potentialité (qui se créé grâce à la caractéristique inhérente au document comme bien d'expérience). [25]

Les notices bibliographiques au service de l'organisation du savoir

Les index bibliographiques sont la clé de ce type de gestion, permettant de reconnaitre tel ou tel ouvrage comme étant Un (une édition particulière dont la matérialité des exemplaires n'importe que peu), ainsi que ses liens avec les précédentes éditions, traductions, changement de format, etc. Ils constituent en eux-mêmes cette notion d'identification, si particulière aux bibliothèques car elle ne nécessite pas une unicité parfaite. Le bien informationnel doit être identifié par son essence, son fond et sa forme, et non par sa matérialité concrète.

---

[23] SAKALAKI, Maria, THEPAUT, Yves. La valeur de l'information, évaluation des biens informationnels versus bien matériels. *Questions de communication*, 2005. N°8. [en ligne] Disponible sur https://questionsdecommunication.revues.org/5300#quotation [consulté le 14/07/2017]

[24] SALAÜN, Jean-Michel. *Les sept piliers de l'économie du document.* 2006. [en ligne] Disponible sur http://blogues.ebsi.umontreal.ca/jms/index.php/post/2006/10/05/86-les-sept-piliers-de-l-economie-du-document [consulté le 18/08/2017]

[25] SALAÜN, Jean-Michel. *Vu Lu Su ; les architectes de l'information face à l'oligopole du web*. Paris : Éditions La Découverte, 2012.

La notion de notice bibliographique*, qui comporte la carte d'identité du livre, est une notion qui intègre déjà cette idée de « donnée sur la donnée », induisant une rétrospection sur soi et sur ses éléments descriptifs les plus pointus. Qu'est-ce qui fait de moi ce que je suis ? Mon identification ne se fait pas par un élément de clé (un numéro unique) mais par l'agrégation de toutes mes caractéristiques combinées. C'est ce qui me différencie de l'autre. Car si je ne suis pas différent de l'autre, l'autre est moi de manière indistincte. Ainsi, de simple « fiches informatives », les notices bibliographiques deviennent des documents en soi. La gestion de ces données cruciales conditionne toute l'organisation de la connaissance acquise et du fonctionnement de la bibliothèque.

Paul Otlet avait bien compris cet enjeu puisque de 1910 à 1934 il met en place le *Mundaneum*, ce réseau de métadonnées manuel présageant étrangement du web futur[26] comportant un ensemble de notices sur des cartes où l'on pouvait faire des requêtes (et recevoir les réponses par courrier postal). Son système, avec une classification à facettes, était déjà conçu dans une optique de globalité, de connaissance universelle, un objectif d'omniscience qui ressort tout à fait dans ses écrits.

> « Dans cette partie on considère l'ensemble de la documentation, ses buts, ses parties composantes, ses unités, ses collections, ses opérations et formats. On examine quelle manière à en former des ensembles et à leur degré supérieur, réaliser une organisation universelle et mondiale. Les moyens, les méthodes, l'outillage, les relations entre organismes sont examinés en fonction de ce but ultime »[27]

Le catalogage et l'indexation sont donc deux opérations au cœur de l'organisation et du fonctionnement des bibliothèques. Dans le monde scientifique, lorsque l'indexation est physique, elle conditionne totalement l'utilisation que l'on va faire de la bibliothèque. Les chercheurs et étudiants qui sont dans des universités de petite envergure sont désavantagés : plus la personne est loin d'une ressource, plus le temps va être long pour qu'elle y ait accès. Avec l'importance que l'on connait du *peer-reviewing*[28] et de la citation entre articles, l'index de citation comme outil constitue une forme ancienne de données liées de bibliothèques (LLD Library Linked Data) qui se visualise par un graphe nodal. Là encore, la gestion numérique qui est faite n'a rien d'une innovation.

La gestion bibliographique en bibliothèque

Dans les années 60, l'IFLA apporte le concept de l'UBC (*Universal Bibliographic Control*) rassemblant de nombreuses caractéristiques déjà existantes sous un même système globalisé. Il s'agit de produire rapidement et de mettre à disposition dans un format adéquat de la donnée bibliographique de base sur toutes les publications au niveau international. Cet idéal transculturel et mondialisé aboutit à la création d'un standard pour la définition des notices afin que celles-ci soient lisibles et intelligibles par des machines. Chaque agence nationale UBC est

[26] La corrélation alla relativement loin puisque Vinton Cerf, l'inventeur du protocole TCP/IP aurait déclaré que l'idée de l'Internet serait née en Belgique. Voir BOULTON, Jim. *The idea of the internet was born in Belgium*. Digital-archaeology.org [en ligne] Disponible sur http://digital-archaeology.org/the-idea-of-the-internet-was-born-in-belgium/ [consulté le 11/08/2017]

[27] OTLET, Paul. *Op. Cit*. p 372

[28] *Peer-reviewing* : forme de validation par les pairs. Lorsqu'un article scientifique est soumis à publication il est lu et étudié par les confrères-consœurs spécialistes dont l'avis en détermine souvent la qualité.

responsable de la création de la notice bibliographique (unique à chaque édition d'un ouvrage) faisant autorité, pour toutes les productions de son pays. Ce beau projet se voit cependant opposer des problématiques dans les années 1970, par l'apparition du besoin émergent de « *prioriser les besoins des utilisateurs nationaux par rapport à l'utilité de conformer à une norme d'uniformisation internationale* ». UBC s'est cependant adapté en proposant des règles relativement souples permettant les variations au niveau national.[29]

Un peu plus tard, dans les années 1980 (et avant l'émergence du web), la production des fiches bibliographiques s'automatise. Les bases de données peuvent être interrogées pour récupérer des notices (comme avec le protocole Z39-50) via des interfaces portail. Le catalogue devient accessible aux utilisateurs et devient un OPAC (*Online Public Access Catalog*). La BnF en fait l'expérience avec le dispositif Telnet. Dans les années 2000, les notices commencent à recevoir des identifiants uniques comme les PPN du Sudoc ou encore, un peu plus tard, des identifiants ARK.[30] Ce ne sont pas uniquement les ouvrages en eux-mêmes qui sont identifiés de manière unique, mais aussi les notices.

Cependant, la bibliothèque voit avec les nouveaux développements technologiques externaliser ses services, qui deviennent à eux seuls des industries culturelles : collections numériques auxquelles elle doit s'abonner (revues scientifiques), recherche d'information (Google), offre directe sur le web de livres numériques ou numérisés. Cette transfusion d'activités d'un acteur à l'autre de la chaîne n'est pas nouvelle, l'imprimerie ayant auparavant renoncé à la fonction d'édition qui était pourtant son exclusivité[31]. En 2004, c'est l'entrée en bourse de Google, avec à peine quelques mois plus tard l'introduction de Google Books. La déclaration faite à ce sujet sera intéressante :

> « La mission de notre compagnie est d'organiser l'information du monde entier et de la rendre universellement utile et accessible. »[32]

Nous avons notre nouveau *Mundaneum*. Les bibliothèques Européennes ripostent avec la création d'Europeana, pour faire face à la menace de l'hégémonie américaine sur le savoir et la culture. Seule la Bibliothèque Municipale de Lyon cède à Google en France. Tout compte fait, la vraie innovation est le web. Les ordres documentaires actuels ne sont en réalité que des réadaptations et réorganisations d'un processus déjà établi. Le web, en apportant cet aspect de diffusion globale des données, met en concurrence les bibliothèques avec des acteurs plus « économiquement agressifs », fournissant du contenu, et les transforme indubitablement en industries de la mémoire. [33]

La cotation dans tout cela reste intimement liée historiquement à la classification, même si le numérique, nous le verrons, tend à bouleverser ces principes et à déplacer les liens surannés entre identification et contextualisation.

---

[29] WILLER, Mirna, DUNSIRE, Gordon. *Op. Cit.*

[30] WENZ, Romain. *Op. Cit.*

[31] SALAÜN, Jean-Michel. *Op. Cit.*

[32] MARTIN, Frédéric. Dans ALIX, Yves (dir.). *Bibliothèques en France 1998-2013*. Editions du cercle de la librairie, 2013. 279p. Chapitre « L'enjeu des bibliothèques et des ressources numériques : la bibliothèque sur la toile ».

[33] *Ibid.*

1.1.2. Dans les archives

D'autres objectifs, d'autres documents

Les archives gèrent des documents en beaucoup plus grand nombre que les bibliothèques, en accroissement constant et exponentiel, et qui sont généralement uniques. L'unité de mesure est l'article, même si cette notion a été longue à mettre en place au vu des débats professionnels sur le sujet ayant eu lieu dans les années 1980.[34] Les modes d'accès à ces documents sont limités par des problématiques de confidentialité, de sécurité, de pérennité. A partir de ce constat-là, les objectifs d'identification ne sont évidemment pas du tout les mêmes que dans les bibliothèques.

Les archives sont soumises à des principes de gestion qui leur sont consubstantiels, et parmi ceux-ci le plus important est celui visant à garantir le respect des fonds. Il contient trois principes fondateurs : respect de la provenance, respect de l'intégrité, respect de l'ordre original. Ces trois principes conditionnent l'organisation complète du classement des fonds, et, parallèlement leur identification.

En effet, comme pour les bibliothèques, la cotation est très liée à la notion de classement. *L'abrégé d'archivistique* la définit comme telle :

> « La cote correspond à la place de l'article dans le cadre de classement ou à sa localisation dans les magasins de conservation » (aussi nommée cadre de cotation).[35]

La définition de la *Pratique archivistique* est quant à elle beaucoup plus développée mais reprend également ces principes : la notion de cote est étroitement liée à la place dans le cadre de classement, à l'adresse dans les magasins et « *tout article doit règlementairement porter un élément d'identification (généralement un ensemble de lettres et de chiffres)*».

Les trois objectifs de la cote

On distingue bien ici une certaine liberté dans l'idée que la cotation puisse être à la fois une adresse (localisation physique, liée au rangement pratique), une identification (dans le sens nominatif, correspondant à l'individualisation de l'article) et une contextualisation (présence logique et hiérarchie intellectuelle dans le fonds). Ces trois aspects se télescopent sur un même outil, la cote, ce qui semble avoir créé des problèmes à l'ensemble de la communauté des archivistes depuis les prémices du métier. Baptiste De Coulon, dans son article *De l'intérêt de la cotation multiple en archivistique* paru en 2016 va jusqu'à la comparaison de la problématique de la cotation à une « boite de Pandore » que l'on préfère tenir fermée. Il parle de réelle « *réticence au sein de la profession à mettre en discussion la cotation […] On toucherait là à un point sensible […]* ».[36] C'est en tout cas quelque chose qui se confirme totalement à l'exploration de la documentation

[34] A titre d'exemple, Gilles Héon publie en 1987 un article intitulé « L'article dans les répertoires : élément de cotation ou élément de rangement ? », auquel Louis Garon répond de manière très incisive dans un article de contre-argumentaire paru en 1989 intitulé « A-t-on besoin d'une nouvelle définition de l'article ? ».

[35] AAF. *Abrégé d'archivistique, Principes et pratiques du métier d'archiviste*. 2012. p 140

[36] DE COULON, Baptiste. *De l'intérêt de la cotation multiple en archivistique*. Site siar.hypotheses.org, 2016. [en ligne] Disponible sur https://siar.hypotheses.org/59 [consulté le 09/05/2017]

professionnelle au sujet de la cotation : les circulaires et décrets s'accumulent se réfutant les uns les autres, *l'Abrégé d'archivistique* de l'AAF et la *Pratique archivistique française* restent vagues sur la question, laissant une grande marge de manœuvre aux décisions et aux organisations individuelles au sein des services d'archives.

En 1987, Gilles Héon fait même référence à la Tour de Babel pour qualifier les problèmes de terminologie au sein de la profession. Pour lui, l'élément existe parce qu'il est identifié. Cet élément est contenu dans un article, qui est une « *unité matérielle d'archives pourvue d'une cote* ». Il met notamment en garde contre l'écueil de confondre les notions de cartons et d'article, et il écrit à ce propos :

> « Deux réalités semblent devoir cohabiter dans cette notion d'article : l'une physique et matérielle (unité matérielle telle qu'elle se présente sur les rayons) et l'autre logique et intellectuelle (unité pourvue d'une cote). Comment, dès lors, une cote, considérée comme permanente et unique, peut-elle être attribuée à une réalité matérielle potentiellement variable tant par sa dimension, à la suite d'une restauration par exemple, que par son rangement dans les magasins. »[37]

De même, Baptiste De Coulon considère que la prolifération des termes de cotation en archivistique est symptomatique du problème de la délimitation de la notion : cadre de classement, plan de classement, code de classification, cote de classement, cadre de cotation, modèle classificatoire…[38] La notion de cote en elle-même est très ambiguë car elle implique les principes logiques, topographiques, ou bien les deux à la fois. L'incapacité à trouver une cohérence dans cette définition de manière internationale, nationale voire au niveau local montre que la maitrise de la question est toute relative.

Pour De Coulon, le regroupement des trois fonctions d'individualisation, de contextualisation et de topographie est la source du problème. Il préconise une séparation de ces différents éléments via la cotation multiple, avec des usages déterminés pour chacune d'elles. La problématique trouve sans doute ses réponses lorsque l'on transpose les concepts à ceux de l'identification numérique : est-il nécessaire d'avoir, comme dans le numérique, une séparation de l'identification unique et pérenne (un identifiant), du contexte évolutif (un ensemble de métadonnées) et de la localisation (accès via un domaine) au sens d'une mécanique d'accès ? En tout cas dans le domaine du numérique, la cotation multiple et la scission de ces éléments permet une indépendance et une souplesse relativement incontournables lorsque l'on gère de la donnée brute. Les cotes physiques uniques entrainent la non-différenciation des formats, il peut donc y avoir des pertes conséquentes de place, ou subséquemment la multiplication des fantômes : c'est un problème que l'on retrouve avec les URL « cassés » sur le web, et bien qu'ils ne soient pas dus aux mêmes causes, ils sont tous autant source de frustration.

Outils et principes d'utilisation de la cote

En ce qui concerne les méthodes de gestion de cote en elles-mêmes dans les archives, les liens entre les différentes entités se font au moyen d'outils papier

---

[37] HEON, Gilles. L'article dans les répertoires : élément de cotation ou élément de rangement ? *La Gazette des archives*, n°136, 1987. p5-16. [en ligne] Disponible sur http://www.persee.fr/doc/gazar_0016-5522_1987_num_136_1_3018 [consulté le 06/07/2017]

[38] DE COULON, Baptiste. *Op. Cit.*

(« *introduction d'un instrument de recherche papier* ») ou informatiques (« *fichiers chaînés, SGBD relationnels,* etc. »). Les instruments de recherche ont une place prépondérante au sein de la gestion d'archives, au même titre que les outils bibliographiques en bibliothèque. La circulaire du 31 décembre 1979 (circulaire de la Saint-Sylvestre ou circulaire W) indique que « *le bordereau de versement est conçu dès l'origine comme un instrument de recherche archivistique, et que les cotes données au moment du versement sont définitives.* » Ils sont donc considérés comme faisant partie des instruments de recherche au même titre que les procès-verbaux de récolement, les états de l'inventaire, les états sommaires, les états des versements, les répertoires (numériques, méthodiques), les index, les guides d'archives, etc.

Les principes de cotation ont évolué par rapport aux contraintes apparues progressivement : l'accroissement de fonds, et surtout la profusion de documents récents qui ne peuvent pas raisonnablement faire l'objet d'une cotation définitive (à cause d'un problème de place et de moyens matériels notamment). L'organisation par série thématiques (et de provenance) qui était en vigueur jusqu'en 1979 (remise en cause avec la circulaire évoquée plus haut) fait place à une cotation continue, où tout appartient à la série W qui était jusque-là inutilisée, ou par millésime. Nous voyons également apparaitre la notion d'identification du service versant et/ou du service d'archives au sein même de la cote, permettant une globalisation des cotes au niveau national. Cette circulaire ne concerne cependant que les archives départementales, mais elles seront bientôt suivies par les archives communales et les archives nationales. Ainsi, les fonds antérieurs au 10 juillet 1940 seront rétrospectivement cotés à la manière ancienne (par série), et la problématique liée au fonds clos/fonds ouvert est conclue par l'institutionnalisation de ces nouvelles méthodes de cotation continue, qui supprime la concordance entre la cote et sa signification méthodique intellectuelle.[39]

Anatomie de la description archivistique

Au-delà de l'article, une gestion de la granularité est faite au sein des classements. L'individualisation de chaque fonds et de chaque série se fait également par une cotation propre. En ce qui concerne la description des fonds, contrairement au domaine de la bibliothéconomie, l'archivistique ne peut pas préconiser systématiquement la reprise tel quel de l'intitulé dans la description et l'attribution du nom. Si pour un ouvrage c'est effectivement l'évidence, pour une archive l'intitulé peut être soit trop redondant, soit trop vague, soit erroné. La description formelle d'un article (l'attribution de « métadonnées ») est normalisée par l'ISAD-G en 1999 et le format EAD (*Encoded Archival Description*) tend à égaliser les pratiques en matière de description des articles pour assurer l'interopérabilité des systèmes. Tous deux proposent des structurations qui permettent de palier plus ou moins bien aux problèmes d'hétérogénéité des cotations.

En outre, De Coulon propose de s'aligner sur les pratiques d'identification numérique afin de s'intégrer aux changements de paradigme, qu'il considère comme étant une « *nécessité pressante à l'heure de la mise en réseau et des mises en ligne de masse.* »[40] Ceci s'accompagne de recommandations pour l'identification, qui sont d'ailleurs très en lien avec ce qui se fait en termes d'identification numérique :

[39] AAF. *Op. Cit.*

[40] DE COULON, Baptiste. *Op. Cit.*

l'identification de l'institution de conservation (que l'on retrouve dans quasiment tous les systèmes d'identification numérique globaux de type DOI, ARK, même ISBN et ISSN), ainsi que la numérotation continue de type *numerus currens*, qui évoque l'opacité des dénominatifs numériques préconisés.

La cotation en archivistique est donc un domaine plutôt obscur dans lequel chaque professionnel évolue avec ses propres convictions. Les différents objectifs de la cote ici ne trouvent pas un équilibre harmonieux. Nous verrons par la suite si l'arrivée du numérique aide à voir plus clairement les possibilités nouvelles de cotation en archivistique.

1.1.3. Dans les musées

Problématiques liées au domaine

Dans les musées, il existe un autre type de confusion terminologique qui concerne les identifiants. La confusion ne se fait pas entre identifier et classer, mais entre identifier et décrire. Les éléments descriptifs de l'objet font partie intégrante de l'identification, ce sont justement ces données-là qui apportent la distinction unique. Les deux types d'identification en pratique, le numéro d'inventaire et la description, ne sont pas forcément liés ou en cohérence l'un avec l'autre. Le recensement attribue une cote (ou numéro d'inventaire) de manière chronologie, au fur et à mesure de l'entrée de l'objet dans la collection « *qu'il sanctionne de manière irréversible* »[41]. La norme ObjectID, censée normaliser la description muséale, ne propose pas un numéro global auquel relier les données (métadonnées*) qu'elle contient, et qui pourrait identifier de manière globale un objet. D'un autre côté, les numéros d'identification (numéros d'inventaires) ne permettent pas de comprendre l'objet (de l'appréhender), et lorsqu'ils sont utilisés ils servent généralement à l'identifier localement.

Plus que des livres ou des documents, le musée gère un patrimoine qui peut prendre toutes les formes, aussi impalpables qu'une performance d'artiste, ou aussi encombrantes qu'un monolithe celte. Cela se divise en plusieurs domaines, puisque l'on dénombre trois types généraux de musée : le musée d'art, le musée de sciences naturelles ou *museum*, et le musée d'histoire.[42] L'objet muséal est, par essence, unique. Contrairement aux bibliothèques ou aux archives, l'objet muséal est tout à fait singulier : il peut appartenir à une collection, un ensemble, voire à une série (artistique, historique…) mais la singularité de chacun des exemplaires en détermine la valeur. De ce fait, il induit encore plus de problématiques de droits de propriété intellectuelle, de droits patrimoniaux, de protection d'objets précieux, et de manière générale une plus grande frilosité au partage en ligne des collections. De plus, les collections sont soumises au principe d'inaliénabilité. « *Une fois entré dans les collections publiques, un objet de musée ne peut pas redevenir « privé ». […] un musée ne peut pas vendre tout ou partie des collections dont il a la charge.* »[43]

André Gob et Noémie Drouguet, dans leur ouvrage sur la muséologie, précisent en outre que « *l'identification des objets et leur conservation s'inscrit dans une démarche scientifique*. » Ici, le terme identification se rapporte à l'appréhension de l'objet, son analyse, et non pas à sa dénomination de manière unique et univoque.

---

[41] GOB, André, DROUGUET, Noémie. *La Muséologie*, Armand Colin, 4e édition, 2014.

[42] *Ibid.*

[43] *Ibid.* p190

Nous voyons bien ici comme la frontière entre les deux notions est facilement franchissable.

Le récolement des collections

Les collections muséales, tout comme les archives, sont soumises régulièrement à des récolements visant à analyser la teneur, la qualité et la complétude des fonds présumés d'une institution. C'est une obligation légale depuis 2002 que d'effectuer de manière systématique et périodique (tous les 10 ans sauf dans des cas particuliers) la vérification des œuvres inscrites, afin de pallier aux inventaires incomplets ou non tenus correctement. Le musée des Beaux-Arts de Lyon, par exemple, a effectué un récolement de 2004 à 2015 sur l'ensemble de ses collections.[44] Les objectifs étaient divers : à la fois des objectifs sécuritaires (vérifier que tout est bien là et dans le cas d'un manquement porter plainte, obtenir une vision globale des collections, restaurer s'il le faut des pièces) ; des objectifs d'amélioration constante du contenu (réévaluer des objets mal identifiés, faire des découvertes en faisant le lien entre différents objets, localiser toutes les pièces) ; mais également des objectifs tournés vers le numérique et l'accès aux collections (compléter les bases de données, mettre en ligne des catalogues, effectuer des numérisations, etc.).

Aux Beaux-Arts de Lyon, un registre général a été tenu jusqu'en 1878 puis ensuite l'organisation a été segmentée en un registre par collection, chacun identifié par une lettre. Les numéros d'inventaires identifiant une ressource étaient constitués de la lettre du dit registre suivi d'une numérotation séquentielle en fonction de l'entrée. En 1935, le registre redevient unique et le numéro d'inventaire se complète de l'année d'acquisition pour faciliter le compte des entrées. En 2002, le musée des Beaux-Arts de Lyon applique le système d'identification préconisé par le Service des Musées de France : l'année d'acquisition, le numéro d'entrée pour l'année, le numéro de l'objet dans l'acquisition même (surtout dans le cas d'une acquisition de plusieurs objets à la fois).[45] Si ce cas peut être anecdotique, nous voyons bien ici que l'identification est systématiquement liée à un registre, un inventaire, elle n'a pas d'existence ni d'utilité propre. Elle ne contient pas non plus d'information de localisation, contrairement aux problématiques rencontrées dans les archives ou dans les bibliothèques. La cote, ou, devrions-nous dire, le numéro d'inventaire, a une fonction d'identification pure, dépouillée de ses fonctionnalités de contexte et de localisation.

La description de l'objet muséal

La création de « métadonnées » pour un objet muséal se fait par la création de fiches, qu'elles soient numériques sur une base de données (base de gestion des collections) ou papier. Chaque fiche comporte le numéro d'inventaire, le titre, les caractéristiques techniques, l'état de conservation et de marquage. A la fin du récolement, le conservateur va valider la conformité de ces données et établir un bilan sous forme de procès-verbal.[46]

---

[44] MUSEE DES BEAUX-ARTS DE LYON. *Le récolement*. Article sur le site du Musée des Beaux-Arts de Lyon, date inconnue [en ligne]. Disponible sur http://www.mba-lyon.fr/mba/sections/fr/collections-musee/vie-des-collections/le-recolement [consulté le 01/06/2017]

[45] *Ibid.*

[46] *Ibid.*

La norme ObjectID, créée par J. Paul Getty Trust et lancée en 1997, est assez parlante concernant la propension des communautés professionnelles de musées à faire un amalgame entre description et identification. Elle nait sous l'impulsion de diverses organisations qui en attendent un bénéfice particulier : des services de renseignements tels que le FBI, Scotland Yard mais également l'UNESCO, Interpol, les musées, etc. Elle est développée par une collaboration étroite entre ces institutions mais également les forces de l'ordre, les experts du marché de l'art et les assurances. Le but premier de la norme est en effet de lutter contre la contrefaçon, le vol, le cambriolage, les demandes de rançons et le trafic illicite d'objets d'art. L'idée est de fournir une documentation de qualité unique : décrire au plus près un objet dans toute son unicité afin de l'identifier.[47] La norme précise également que ces données doivent être conservées dans un lieu sûr (non altérables, non diffusables).

Le catalogue

L'inventaire est à distinguer du catalogue, qui lui est un outil de communication intéressant pour les musées permettant un rassemblement des informations concernant une pièce avec son identification. Il comporte donc, en général, l'identification de l'objet (numéro d'inventaire ainsi que les éléments de description, normés ou pas), les éléments de classement logique et matériel, l'origine et le statut juridique de l'objet, la date d'entrée, l'état de conservation, ainsi que le prix d'achat ou la valeur d'assurance. Contrairement aux bibliothèques et aux archives, ici la problématique n'est pas centrée sur l'organisation en typologies ou nomenclatures : la taxonomie est un outil simple sur laquelle prévalent souvent les principes de collections.

Il y a différents types de catalogues tels que le catalogue raisonné qui porte sur un ensemble homogène d'objet, ou encore le catalogue par ordre d'acquisition. Le musée peut également avoir une activité éditoriale riche : bulletins, annuaire, revues, publications de recherche… Les supports aussi peuvent être variés : catalogues en ligne, catalogues papiers, catalogue sur CD-Rom, etc. Cette production jongle avec les informations contenues de manière à diffuser uniquement ce qui est nécessaire, et peut être sujette à l'attribution d'un numéro ISSN ou ISBN.

La standardisation reste cependant délicate :

> « La très grande diversité des musées et de l'apparition de très nombreuses bases de données disparates, utilisables seulement à des fins de gestion des collections au sein de l'institution muséale et le plus souvent incompatibles entre elles. L'extrême variété des collections et la grande difficulté de les cataloguer sous un canevas descriptif commun rendent d'autant plus ardue la recherche d'une certaine standardisation. »[48]

Le musée aussi doit s'organiser autour de problématiques de cotation complexes qui peinent à être résolues par des solutions applicables globalement. Les efforts de normalisation ont permis de soulever d'autres questionnements encore sur la manière de coter et de décrire afin de satisfaire au mieux aux exigences propres à ces institutions.

---

[47] ICOM, *Norme ObjectID*, Site Archives ICOM Muséum.fr [en ligne] disponible sur : http://archives.icom.museum/objectid/how_fr.html [consulté le 27/04/2017]

[48] GOB, André, DROUGUET, Noémie. *Op. Cit*.

1.1.4. Dans les entreprises et fournisseurs de contenu

Les éditeurs à la proue de l'identification normalisée

L'identification des documents « produits », tels que les livres, les journaux, les revues, etc. a été une priorité pour les éditeurs. En effet, le modèle économique dépendait de leur capacité à gérer globalement leurs collections, leur stock, et leurs systèmes. L'identification normalisée permettait de leur offrir une « *gamme d'utilisation, d'échange de données, d'interrogation* » de laquelle découlaient la mise à jour instantanée des catalogues, l'identification sûre des commandes, le traitement automatique, etc.[49] En bref, une gestion permettant de leur assurer une compétitivité économique inégalée jusqu'à présent. Cliff Morgan, de John Wiley & Sons, considère même que ce sont les éditeurs qui ont le plus travaillé à développer les identifiants pérennes au vu du gain financier que cela représentait pour eux. Ils ont travaillé plus en profondeur que les autres communautés professionnelles, notamment sur les fonctionnalités, les exigences des utilisateurs et les standards d'identifiants pérennes.[50] C'est une constatation que l'on peut faire en effet en regardant l'origine des identifiants actuels : ils se basent sur des principes institués non pas par les bibliothèques, les archives ou les musées, (qui, nous l'avons vu, ont des difficultés à appliquer des systèmes de cotation interopérables et efficaces en toute circonstances), mais par des organisations à but lucratif.

Les éditeurs à l'origine de l'ISBN et de l'ISSN

Nous avons des exemples assez classiques de normalisation dans le domaine du livre et des publications sérielles : l'ISBN et l'ISSN. La création de l'ISBN, *International Standard Book Number*, originellement SBN (pas encore internationalisé) s'est vue impulsée par un libraire-distributeur anglais, W. H. Smith qui en 1965 souhaitait automatiser ses systèmes et délocaliser son stock. Jusqu'à lors, chaque éditeur possédait ses propres identifiants « maison » et l'idée d'une normalisation n'était pas encore au goût du jour. L'association des éditeurs (*Publishers Association*) ayant pris en main l'idée, le concept fédéra ensuite nombre d'acteurs importants tels que la British National Bibliography qui souhaitait également automatiser, ou encore le Greater London Council. Les experts s'étant penchés sur la question, en 1967 naquit le SBN. Deux ans plus tard, le standard ajoute un chiffre à son identifiant originel de 9 chiffres, qui devient, sous l'égide du comité ISO/TC 46, le numéro normalisé international du livre.[51]

Ce schéma, proposant un système ingénieux permettant une souplesse très innovante, permet de s'adapter à des tailles d'entreprises très diverses : de quelques publications à de très grosses productions au-delà de 20 000 titres annuellement. Il comporte plusieurs « tiroirs » d'identification : l'identification globale du secteur mondial (pays ou zone), une identification distincte pour chaque fournisseur de contenu (plus le chiffre est grand, moins l'organisation produit du contenu), et

[49] HONORE, Suzanne. *La numérotation normalisée internationale du livre (International Standard Book Number.)* Site du Bulletin des Bibliothèques de France [en ligne] Disponible sur http://bbf.enssib.fr/consulter/bbf-1969-08-0321-001 [consulté le 27/04/2017]

[50] ERPA Seminar. *Persistent Identifiers, Final Report*. Cork, Ireland. 17-18 June 2004. Séminaire Erpanet, [en ligne] Disponible sur http://www.erpanet.org/events/2004/cork/Cork%20Report.pdf [consulté le 12/04/2017]

[51] HONORE, Suzanne. *Op. Cit.*

ensuite une identification de la production sur une plage de chiffres laissée à la pratique interne d'attribution de chaque organisation.

L'ISSN est également un exemple intéressant d'association des logiques éditeurs et bibliothèques : l'influence initiale est celle des éditeurs, mais le réseau d'attribution est un réseau de bibliothèques (ce qui explique pourquoi il y a un registre alors que l'ISBN n'en a pas). C'est l'UNESCO, qui en 1969 a développé le principe de l'ISSN pour la première fois dans le cadre du programme UNISIST, un projet ayant pour but la création d'un registre de publications en série de l'information scientifique mondiale. Alors appelé ISDS (*International Serials Data System*), il était géré par un Centre international d'enregistrement (le CIEPS, Centre International d'Enregistrement des Publications en Série) sous l'égide de l'UNESCO et du gouvernement français[52]. Celui-ci coordonnait également des centres nationaux chargés de procéder à l'enregistrement des numéros sur leurs zones respectives. Ces centres et le CIEPS sont encore actifs pour la gérance de l'ISSN que l'on connait actuellement, qui a été formalisé en 1975 dans la norme ISO 3297 encore utilisée et mise à jour régulièrement[53].

Les éditeurs ont en outre un rôle majeur à jouer dans cette attribution car il leur incombe de maintenir le lien de manière permanente entre le numéro ISSN sur leurs publications, dans leurs catalogues et sur leurs différents moyens de communication.

Rendre à César ce qui appartient à César

La question des identifiants dans les entreprises qui fournissent du contenu informationnel est aussi très liée aux droits, notamment dans le domaine de la musique et de l'audiovisuel en général où il y a énormément de droits différents à prendre en compte et qui se chevauchent. La performance, l'enregistrement, la composition, la diffusion… les acteurs du domaine ont beaucoup à gagner à développer ensemble des standards et des accords.[54] Cela concerne les droits de propriété intellectuelle en général, les droits moraux et patrimoniaux, la reconnaissance d'une même entité sous différents patronymes, la gestion du *versioning* d'une œuvre, etc. Jusqu'aux années 2000, cet aspect était traité localement et individuellement, comme chez les éditeurs pour le livre avant 1969. Actuellement, il existe des bases d'identification standardisées et internationalisées pour la gestion de ce type de droits tels que l'IPI, l'ISNI ou encore l'IPN qui, à termes, devraient se rejoindre.[55]

Pour les livres, la transposition de l'ISBN du monde physique au monde numérique pose un problème : si dans le monde physique l'enjeu est surtout d'identifier la source du produit (ce que l'ISBN fait très bien du reste), dans le monde numérique il faudrait également pouvoir indiquer des droits sous-jacents au contenu. En effet, les ayant-droit n'ont peut-être pas de droits sur les contenus électroniques, les droits ont peut-être été transférés entre temps, ou bien le contenu n'est plus lié, etc. L'ISBN n'est qu'un identifiant auquel aucune métadonnée n'est

[52] BOSSUAT, Marie-Louise, GIRARD, Christine. Le numéro international normalisé des publications en série (ISSN). *Bulletin des Bibliothèques de France*. 1974. N°12.

[53] ISSN. *Le fonctionnement de l'ISSN*. Site ISSN.org [en ligne] Disponible sur http://www.issn.org/fr/le-centre-et-le-reseau/notre-organisation/les-statuts/ [consulté le 16/08/2017]

[54] GREEN, Brian, BIDE, Mark. *Unique Identifiers: a brief introduction*. 1999. 11p. [en ligne] Disponible sur http://www.bic.org.uk/files/pdfs/uniquid.pdf [consulté le 29/11/2016]

[55] Nous développerons ces aspects dans la deuxième partie de ce mémoire.

rattachée, là encore il est dépouillé de ses informations de contexte et de localisation.[56]

Dans le cas d'un produit pur, informationnel ou non, l'identification par code-barres est utilisée dans les entreprises afin de garantir une traçabilité des biens et des objets de consommation. Cela permet notamment l'amélioration de l'efficacité de la gestion logistique, le contrôle, le suivi des stocks, le suivi jusqu'à la réception de la marchandise, le respect de la règlementation, la garantie de la qualité sanitaire, etc.[57] Il s'agit bien entendu d'une identification sérielle, correspondant à un modèle de produit bien défini mais dont on ne va pas prendre en compte l'unicité en tant qu'objet. L'ISBN et l'ISSN sont d'ailleurs réutilisés dans les codes-barres EAN normalisés, car les publications sont des produits avant tout. L'utilisation courante du code-barres intervient dans les années 1970, et se développe jusqu'à devenir universelle grâce à la simplicité du système. C'est un exemple d'identification qui s'est distinguée d'elle-même pour être adoptée comme norme.

Conclusion de la partie 1.1

En conclusion de ce premier bilan global de l'apparition et du développement des identifiants et des métadonnées dans les différents secteurs, nous pouvons constater que chaque domaine s'est plus ou moins développé selon ses propres contraintes et objectifs, sans vraiment chercher à s'appuyer sur les retours d'expérience des uns et des autres. Nous constatons beaucoup d'initiatives individuelles qui remplissent des fonctions sans convenir à d'autres, et qui sont difficilement interopérables. La problématique de « globalité » et de socle commun d'échange est mise à mal malgré quelques efforts manifestes pour trouver des terrains d'accord commun, notamment entre éditeurs et bibliothèques, (qui du reste travaillaient sur les mêmes objets, ce qui pourrait être une explication). En termes de palmarès, il semblerait que les fournisseurs de contenus et les entreprises soient plus avancés globalement que les institutions culturelles classiques puisqu'elles bénéficient d'une motivation pécuniaire forte. Parmi les institutions publiques, la meilleure organisation et la plus grande implication dans ces questions revient aux bibliothèques qui cherchent à tout prix à s'intégrer dans les nouveaux paradigmes numériques et à étendre leur domaine de compétence. Les archives et musées sont tout autant en difficulté avec ces concepts d'identification, bien qu'ils aient des gestions totalement différentes en soi. Les archives se tournent plus facilement vers les nouvelles problématiques du numérique avec des projets de dématérialisation intéressants ; les musées quant à eux sont sur un repli timide vis-à-vis des politiques d'*open-access* et de diffusion, bien qu'ils soient intéressés et innovants numériquement pour ce qui est de la médiation.

L'identification est donc une notion très floue dans ces domaines, elle s'est construite progressivement et par tâtonnements. Les trois aspects clés de l'identification : singulariser, contextualiser, localiser ont été chacun traités différemment, parfois mélangés, parfois segmentés, parfois carrément supprimés, il en résulte une hétérogénéité globale dans l'identification des contenus dans les institutions. Il est intéressant maintenant de voir ce que le numérique apporte

[56] *Ibid*.

[57] BELCADHI, Feriel. *Installer un système de traçabilité au sein de votre entreprise*. Site Usine nouvelle.com, 2015. [en ligne] Disponible sur : http://www.usinenouvelle.com/expo/guides-d-achat/installer-un-systeme-de-tracabilite-au-sein-de-votre-entreprise-18 [consulté le 15/07/0217]

concrètement et comment chacune de ces institutions a su en tirer le meilleur parti pour ses objectifs propres.

1.2. PRINCIPES APPORTES PAR LE NUMERIQUE

1.2.1. Ressources, web document et objets réels. Qu'identifie-t-on ?

La notion de « sens » dans l'attribution des noms

« Ceci n'est pas une pipe »

En 1927, Magritte illustre le paradoxe de l'identification avec son tableau de la pipe. Il met en perspective les différents niveaux de sens que l'on peut percevoir : le sens concret, le tableau en lui-même ; le sens abstrait, l'idée que l'on se fait d'une pipe en tant qu'objet ; et enfin le nom que l'on donne, l'idée évoquée par le langage, le mot « pipe ».

Tim Berners-Lee montre dans les annexes de son document *What do http URI identify* une photo de l'œuvre de Magritte et demande au lecteur de choisir, « *est-ce :*

1.Une pipe

2.Je ne sais pas, mais certainement pas une pipe

3.Une contradiction

4.Une peinture de Magritte

5.Une photographie d'une peinture de Magritte

6.Une représentation d'une photographie d'une peinture faite de 341 632 bits

7.Les réponses 4, 5, et 6 mais certainement pas la réponse 1 »

La solution proposée en 7 est intéressante : « *Réponse 7. Notez ici que le web tolère la souplesse en ce qui concerne les différentes représentations d'une même image, mais pas sur le niveau sémantique entre une image et un objet réel* ».[58] L'on distingue là une notion qui touche au « sens », à la sémantique, la signification d'un signe, qu'il soit visuel (gestuel, pictural, écrit), ou oral (parlé).

Il y a deux visions qui s'opposent lorsqu'il s'agit de trouver du sens « au sens ». Catherine Legg nous en apprend plus dans l'exposition de deux approches différentes : la vision Cartésienne et la vision Piercienne. L'idée clé de la première est que le sens d'un signe est l'intention de son producteur. C'est privé, lié à l'esprit de quelqu'un, et c'est indiscutable, l'opinion de la personne qui l'a produit fait autorité. Cela correspond à la notion d'idée de Descartes : « *L'erreur est possible, mais pas en ce qui concerne le sens des idées de quelqu'un, seulement sur la forme avec laquelle l'esprit se représenta la réalité.* »[59] La vision Piercienne, qui émane de Charles Sanders Pierce, fondateur du pragmatisme, en est une alternative qui se veut moins radicale et moins individualiste. Pour lui, le sens d'un signe est le procédé même d'interprétation quand le signe est évoqué. C'est la relation irréductible entre trois éléments :

- Le *Representatem* (la représentation, le mot)

[58] BERNERS-LEE, Tim. *What do HTTP URIs Identify?* W3C, 2002. [en ligne] Disponible sur https://www.w3.org/DesignIssues/HTTP-URI.html [consulté le 31/03/2017]

[59] LEGG, Catherine. *Pragmaticsm on the semantic web*. 2010. [en ligne] Disponible sur http://www.nordprag.org/nsp/1/Legg.pdf [consulté le 01/06/2017]

- L'objet (l'objet en lui-même, la réalité)
- L'*Interpretant* (les utilisations qui sont faites de ce mot)

Cela correspond à la notion de communication et de réutilisation : le mot ne sert à rien s'il n'est pas partagé, réutilisé. On peut lui ajouter du sens, voir modifier ce sens à chaque nouvelle utilisation. De plus, les mots et signes utilisés évoluent parfois en dehors de l'esprit de leur producteur. Le sens n'est plus un objet, c'est un processus. Dans la pensée Cartésienne*, pour savoir ce que veut vraiment dire un signe il faut être dans la tête de celui qui le pense. Dans celle de Pierce*, il faut analyser l'utilisation qui est fait de ce signe. La responsabilité du « sens » est déplacée à la communauté qui le pratique.[60]

Paul Otlet, quant à lui, définit 4 niveaux de sens : 1. La réalité (exemple : une science, un paysage, une personne), 2. L'image qui reproduit la réalité, 3. La reproduction d'une reproduction de la réalité, 4. Les écrits, qui peuvent être relatifs directement à la réalité ou bien à une image et, dans ce cas, ils sont relatifs à une reproduction de la réalité, ou relatifs à une reproduction d'une reproduction de la réalité (un écrit concernant un tableau, etc.)[61]

> « Le signe d'un objet amène à une interprétation (l'interpretant) qui elle-même comme signe peut mener à d'autres interprétations. »[62]

Cette idée appliquée aux identifiants montre que le signe identifiant se doit d'être compris comme un index : il pointe sur une représentation de quelque forme que ce soit, et cette représentation a une relation avec le signe. Bergman pose la question suivante : l'image d'un toucan correspond-t-elle à un toucan en particulier, un individu spécifique avec un sexe, un âge, parmi des milliers d'individus spécifiques d'une famille de toucan, qui elle-même fait partie de 40 races de toucan ?[63] Il y a de multiples niveaux de compréhension de la ressource : il faut toujours un contexte et une utilisation pour qualifier quelque chose, c'est le propre du langage.

Le problème http Range 14

David Booth sépare de manière radicale les choses identifiées sur le web : celles qui existent sur le web, et celles qui n'existent pas sur le web (les objets physiques et les objets abstraits). Les choses qui existent sur le web sont a priori déjà identifiées par des URL.[64]

---

[60] Nous verrons en partie trois ce qu'impliquent ces schéma lorsqu'on les applique au web sémantique.

[61] OTLET, Paul. *Op. Cit.*

[62] BERGMAN, Mike. *Give me a sign: what do things mean on the semantic web?* Mkbergman.com, 2012. [en ligne] Disponible sur: http://www.mkbergman.com/994/give-me-a-sign-what-do-things-mean-on-the-semantic-web/ [consulté le 18/05/2017]

[63] *Ibid.*

[64] BOOTH, David. *Op. Cit.*

C'est à ce point névralgique qu'apparait le fameux problème *http range 14**, crise identitaire qui alimente les débats depuis des dizaines d'années dans la communauté web. Quand il s'agit de désigner par un identifiant une description d'un objet physique, est-ce que celui-ci identifie l'objet en lui-même ou est-ce qu'il identifie la description ? Cette question est source d'ambiguïté, très néfaste dans le monde du web et des machines, qui ne savent pas prendre en compte le contexte. Comme l'explique Booth, si quelqu'un évoque l'URL « http://exemple.org/amour », comment savoir si celle-ci se réfère au concept de l'amour, une localisation sur le web, un document particulier ou le nom lui-même ? Nous retrouvons d'ailleurs parmi eux nos 4 concepts évoqués dans la première partie :

- La dénomination en elle-même (l'identification, le *representant*)
- L'objet réel (le concept)
- La localisation (une adresse sur le web)
- Le document en particulier, peut-être une version susceptible d'être actualisée ou un document précis individuellement identifié (*l'interpretant*)

En informatique, le concept de l'UNA (*Unique Name Assumption*) fonctionne selon le principe suivant : si deux objets ont le même identifiant, alors ils sont un seul et même objet. Inversement, deux objets qui n'ont pas le même identifiant sont différents de fait. Or, selon la loi de l'identité de Leibniz, la notion d'identique n'a pas de sens. Lugwig Wittgenstein, philosophe et mathématicien autrichien, la résume ainsi : « […] *dire que deux choses sont identiques n'a pas de sens, et dire qu'une chose est identique à elle-même c'est ne rien dire du tout*. »[65] Mais dans ce cas, pour les différencier, il faudrait obligatoirement un identifiant différent pour chacun de ces 4 éléments. Les différentes solutions proposées au *http range 14* n'ont pas été viables et n'ont finalement pas pris racine. Le problème s'appelle désormais *Issue 57*. Nous développerons ces questionnements dans la partie trois de ce mémoire.

Terminologie et concepts

L'ambiguïté de la terminologie « ressource, document, web document, objet documentaire » dans la langue n'arrange rien. La notion de *web document* ne doit pas être confondue avec la notion de fichier. Il s'agit d'entité, comme une page. Un *web document* existe si toutes ses caractéristiques essentielles peuvent être rassemblées et transmises dans un message, et s'il peut être représenté sur le web[66]. Dans le web 1.0, le web classique, les identifiants étaient utilisés pour les *web document* et la notion de l'identité même de la ressource n'avait pas trop d'importance. Actuellement, les objectifs du web sémantique qui consistent à faire des déclarations sur des éléments non localisés sur le web apportent de nouvelles problématiques.

Il y a aussi une différence entre un système de fichier, où chaque identifiant correspond toujours à une représentation spécifique (le fichier dans un format), et le web, où chaque identifiant correspond à une entité conceptuelle (la ressource) qui peut alors avoir plusieurs représentations suivant les capacités de lecture du

[65] Ludwig Wittgenstein, cité dans HYVONEN, Eero. *Publishing and Using Cultural Heritage Linked Data on the Semantic Web*. 2012

[66] AYERS, Danny, VÖLKEL, Max. *Cool URIs for the Semantic Web*. W3C Interest Group Note, 2008. 16 p. [en ligne] Disponible sur https://www.w3.org/TR/cooluris/ [consulté le 29/11/2016]

navigateur (selon des mécanismes de négociation de contenu dont nous parlerons plus tard).

L'on pourrait également faire une distinction entre des ressources informationnelles (dont la vocation est de transmettre des informations, définition correspondant logiquement à un *web document* par exemple), et des ressources non-informationnelles, c'est-à-dire des objets réels qu'ils soient abstraits ou concrets (des toucans et des planètes).[67] Pour Bergman, la distinction « magique » entre ressource informationnelle et ressource non-informationnelle est un effet collatéral de la logique Cartésienne, qui veut que ce soit le producteur d'un objet qui soit la seule autorité capable de dire ce qu'elle est. Qu'est-ce qui crée réellement la limite entre l'un ou l'autre ? Nous pouvons être nous-mêmes des ressources informationnelles si nous partons du principe que tout ce que l'on voit est une donnée traitée par notre cerveau, une information. L'identité d'une chose n'est donc jamais absolue mais contextuelle (ou fonctionnelle). Et le problème vient encore une fois du fait que les machines ne saisissent pas réellement le contexte.

Ce qui est certain, pourtant, c'est que tout le monde s'accorde sur la notion de référent*. Chaque entité qui a besoin d'être reconnue distinctement dans le réseau devrait recevoir au moins un identifiant public pérenne. En étant assignée à cet identifiant elle devient son référent. Ce référent (cette entité dont nous parlons) peut être de n'importe qu'elle forme, mais le lien entre l'identifiant et son référent doit être explicite, et l'on doit pouvoir utiliser une technologie permettant d'obtenir : soit un objet numérique qui serait la représentation d'un objet réel, soit un objet numérique qui apporterait suffisamment d'informations pour constituer une représentation.[68]

Nous retiendrons donc que l'on peut identifier :

- des objets documentaires (ou *web document*) qui sont déjà identifiés sur le web, mais que l'on cherchera à identifier de manière pérenne ;
- des représentations virtuelles d'entités du monde réel, mais cela pose le problème de la cible de l'identification : l'objet, la représentation, ou les deux ? (*http range 14*, ou *Issue 57*).

Ces entités pourront avoir plusieurs identifiants, qui correspondront aux différentes représentations que l'on peut trouver d'eux et qui seront leurs référents.

1.2.2. URI, http URI, URL, URN, IRI… Quelles différences ?

Distinguer les principes liés aux schémas d'identification

A présent, rentrons dans les notions fondamentales qui nous occupent : les schémas d'identifiants principaux sur le web et dans le numérique, l'URL (*Uniform Resource Locator*) et l'URI (*Uniform Resource Identifier*). L'URN (*Uniform Resource Name*) est une troisième notion qui se greffe aux premières car elle est

[67] C'est notamment le point de vue de Ian Davis dans son article *Is 303 really necessary ?* publié en 2010. Pour lui, la seule façon de le savoir est de le déréférencer. Si le statut code est 200, alors il s'agit d'une ressource informationnelle.

[68] PASKIN, Norman, RUST, Godfrey. *Principles of identification*. Linked Content Coalition, 2014. [en ligne] Disponible sur http://www.linkedcontentcoalition.org/phocadownload/principles_of_identification/LCC%20Principles%20of%20Identification%20v1.1.pdf [consulté le 29/11/2016]

intéressante à relier avec les principes abordés précédemment, mais ne se trouve pas réellement au même niveau conceptuel. Nous les distinguerons ainsi :

L'URL est liée au principe de la localisation, à une adresse : c'est une « *chaîne de caractères permettant de localiser une ressource sur le web. Cette chaîne est précédée du préfixe correspondant à la localisation des documents sur le web.* » [69] Plus qu'une adresse, c'est même une méthode d'accès, elle est composée d'un préfixe actionnant le protocole responsable de la requête de l'information.

L'URN est liée au principe de la dénomination, l'identification au sens propre : c'est une « *chaîne de caractères permettant d'identifier (par son nom) une ressource sur le web. Cette chaîne est précédée du préfixe urn.* »[70] Elle n'est pas en soi « actionnable », dans le sens où la taper dans la barre de requête d'un navigateur ne permettrait pas d'accéder à la ressource.

Enfin, l'URI est un principe pouvant regrouper l'une ou l'autre de ces notions, voire les deux : c'est une « *chaîne de caractères permettant d'identifier une ressource sur le web par sa localisation ou par son nom. Cette chaîne est précédée d'un préfixe enregistré tel que http, urn, doi, ark…* » [71].

L'on pourrait dire que l'URL et l'URN sont des sous-ensembles de l'URI, même si actuellement la spécification URI demande à ce que ces deux termes ne soient plus employés (cela entraine cependant des problématiques que nous ne traiterons pas ici, c'est pourquoi nous emploierons tout de même des deux termes).

L'on parle d'identifiant actionnable* lorsque celui-ci permet d'accéder directement à la ressource qu'il identifie sur le web. Dans beaucoup de cas, le déréférencement*, (ou l'action d'actionner l'URI), ne peut se faire directement via un navigateur. Les URI qui commencent par des noms plutôt que par des protocoles doivent être déréférencés via des serveurs spécifiques à leur contexte.

Uniform Resource Identifier

L'URI, littéralement identifiant uniforme de ressource, correspond à une syntaxe d'identifiant normalisée par le W3C (*World Wide Web Consortium*) dans la recommandation RFC 3986 *Uniform Resource Identifier (URI) : Generic Syntax*. La spécification URI définit une implémentation pour accéder à un lien sur un serveur de fichier (souvent grâce au protocole http mais d'autres sont autorisés), et une syntaxe pour référencer.[72]

Cette syntaxe est la suivante :

- Un préfixe qui donne le contexte que l'on va nommer *scheme,* qui sera soit un protocole permettant à l'URI d'être actionnable (http, mailto, ftp, telnet, file…), soit un type d'identifiant de nom (urn, doi…) « *Chaque URI*

[69] Ministère de la Culture et de la Communication. *Identifiants pérennes pour les ressources culturelles ; Vade-mecum pour les producteurs de données*. Version 1.0. 2015. [en ligne] www.bnf.fr/documents/identifiants_perennes_vademecum.pdf [consulté le 29/11/2016]

[70] *Ibid.*

[71] *Ibid.*

[72] PASKIN, Norman, RUST, Godfrey. *Op. Cit.*

commence par un nom de schéma qui se réfère à une spécification pour l'allocation des identifiants au sein de ce schéma. »[73]

- Un élément qui désigne l'autorité nommante: un nom de domaine dans le cas d'un déréférencement par un navigateur, ou encore un identifiant de cette même autorité au sein d'un système (NAAN : *Name Assigning Authority Number*). Cela permet la globalité du système et son intégration dans le web.
- Le nom concret de la ressource (chaîne de caractères qui identifient cette ressource de manière unique au sein de ce système et pour cette autorité)[74].

Tout identifiant respectant cette syntaxe est une URI. Les *URI http* sont des URI qui sont déréférençables via un navigateur web. Ils sont d'ailleurs vivement recommandés par le W3C, qui considère que leur utilisation serait un moyen d'uniformiser et d'assurer l'interopérabilité des systèmes.[75]

Deux variantes aux URI peuvent se trouver mais sont assez spécialisées. La première correspond aux abréviations QName non standardisées qui servent à en faire un « URI Compact » (CURIE), utilisé en informatique dans les applications locales. Quand celles-ci sont liées au web, un système permet d'automatiquement les traduire en URI http avec le nom de domaine adéquat.[76] La deuxième est l'émergente IRI (*International Resource Name*), basé sur le principe que l'URI est trop restrictif au monde occidental puisqu'il n'autorise la construction de l'identifiant que sur une partie bien délimitée de l'ASCII. (Par exemple, les caractères spéciaux doivent être encodés grâce à des pourcentages pour être pris en compte[77]). L'IRI prend en compte l'Unicode ISO 10646.[78]

Uniform Resource Locator

L'URL, comme nous l'avons vu, est l'identifiant de déréférencement par essence. Il utilise le *scheme** http et identifie une ressource principalement par le mécanisme qui permet d'y accéder.[79] Sa construction se fait en deux parties : méthode://localisation. Cette localisation permet en premier lieu la compréhension par la machine d'une série de paramètres qui appellent une base de données, puis d'une chaîne de caractères qui va être interprétée grâce à un annuaire pour finalement donner accès à la ressource. Les URL utilisent de ce fait le DNS qui va les « résoudre » de cette manière par une requête. Pour résumer, l'URL identifie ce qui existe sur le web, l'URI identifie sur le web tout ce qui existe.[80]

---

[73] "*Each URI begins with a scheme name, as defined in Section 3.1, that refers to a specification for assigning identifiers within that scheme*. » BERNERS-LEE, Tim. *Uniform Resource Identifier (URI): Generic Syntax.* The Internet Society, 2005. [en ligne] Disponible sur http://www.ietf.org/rfc/rfc3986.txt [consulté le 12/04/2017]

[74] ARCHIMBAUD, Jean-Luc. *Identifiants des documents numériques : ISBN, ISSN, URL, Handle, DOI, OpenURL…* 2015. [en ligne] Disponible sur http://archivesic.ccsd.cnrs.fr/sic_01068135/ [consulté le 29/11/2016]

[75] L'utilisation d'URI http est une condition importante de l'attribution des 5 étoiles de qualité du Linked Open Data, que nous verrons par la suite.

[76] WILLER, Mirna, DUNSIRE, Gordon. *Op. Cit.*

[77] A titre d'exemple, le point d'interrogation est encodé « %3F ».

[78] HYVONEN, Eero. *Op. Cit.*

[79] BERMÈS, Emmanuelle, ISAAC, Antoine, POUPEAU, Gautier. *Op. Cit.*

[80] *Web sémantique et web de données, Sensibilisation à l'évolution des catalogues*. Programme Transition bibliographique, Réseau national des formateurs, 2016. [en ligne] Disponible sur https://www.transition-bibliographique.fr/wp-content/uploads/2016/04/Web_de_Donnees_26-02-2016_Version_Courte.pdf [consulté le 16/07/2017]

Sur le web, le nom de domaine*, comme nous l'avons évoqué dans notre introduction, est un lieu virtuel (une adresse logique plus que physique) correspondant en réalité à des lieus de stockage divers sur des serveurs pouvant être appelé en fonction de la disponibilité. En France, le nom de domaine est délivré par l'AFNIC (Association Française pour le nommage Internet et Coopération) qui gère et assure la gestion des extensions françaises[81]. A l'international, c'est l'ICANN (*Internet Corporation for Assigned Names and Numbers*) qui gère les *top level domains*[82].

Il est très important de comprendre une chose à propos des URL : ce sont des identifiants à part entière, et non des chemins. Notre mode de pensée calibré « Windows », « système de fichiers » pourrait imaginer que l'URL suivant « http://www.exemple.fr/site-interessant/pages/fichier-interessant.pdf » mène à un fichier PDF nommé « fichier-interessant », qui serait contenu dans un dossier « pages », lui-même contenu dans un dossier « site-interessant », dossier lui-même hébergé quelque part sur un serveur correspondant au nom de domaine « www.exemple.fr ». En réalité, cet ensemble de caractères pourrait tout à fait mener à tout autre chose sur le site « www.exemple.fr ». Tout ce qui appartient au suffixe de l'URL à partir de la fin du nom de domaine est un identifiant qui peut être signifiant (compréhensible par l'œil humain) mais qui ne garantit aucunement son contenu.

Là réside tout le problème des URL, qui peuvent très bien induire en erreur de par leur signifiance sémantique ou logique, tout en pointant sur des ressources différentes et/ou cassées, menant à une réponse code statut 404*. Cela mène la communauté web à une méfiance justifiée pour les URL qui sont vues comme des adresses non fiables. En réalité, il serait tout à fait possible d'instaurer des URL pérennes qui seraient utilisables et fiables, nous verrons cela en troisième partie de ce mémoire.

Uniform Resource Name

Le système URN est un peu particulier car il est géré et contrôlé (par l'IANA, *Internet Assigned Numbers Authority*, un département de l'ICANN), contrairement aux URI et aux URL qui sont, de fait, déclarés comme tels à partir du moment où l'identifiant en question correspond à leur syntaxe. Il consiste en l'attribution d'un nom unique et univoque pour une ressource, pérenne, qui identifie une entité sans donner aucune indication sur le moyen d'y accéder. Une organisation quelle qu'elle soit qui produit des données et des documents peut faire une demande pour être enregistrée par l'IANA comme pouvant attribuer des URN à ses ressources.

Actuellement, il existe 62 systèmes d'identifiants qui sont enregistrés en tant qu'URN, parmi eux isan, isbn, issn, ietf, uuid, nfc... Ces acronymes représentent souvent la structure émettrice mais constituent l'identifiant d'espace de nom* (*namespace identifier*) qui va permettre la contextualisation de la chaîne de caractères qui suit et qui identifie la ressource.

La syntaxe sera ainsi dénotée : « urn:isbn:123456789 ». Comme nous le constatons, ce bout simple ne permet pas le dérérencement. De ce fait, la plupart des implémentations URN « embarquent » l'URN dans un URI http, qui sera construit

[81] *Missions de l'Afnic et axes stratégiques 2017-2019*. AFNIC [en ligne] Disponible sur https://www.afnic.fr/fr/l-afnic-en-bref/presentation/missions-de-l-afnic-et-axes-strategiques-2017-2019-6.html [consulté le 11/08/2017]

[82] *Top level domain*, ou TLD correspond à des domaines de premiers niveaux, par exemple le .fr, .org, .com, etc.

par des serveurs proxy dits « résolveurs ». Malheureusement, il n'existe pas de service global permettant la résolution d'URN, leur utilisation pratique reste de ce fait relativement limitée.

Cecil Somerton, du *Treasury Board of the Government of Canada* pose une question intéressante : qu'est-ce que les utilisateurs veulent réellement voir rester pérenne ? Est-ce l'accès à l'information elle-même, la localisation de l'information concrète ou uniquement son identité ?[83]

1.2.3. Le web sémantique, le web de données et les données liées

Qu'est-ce que le web sémantique ?

Le web sémantique et le web de données sont des notions qui se superposent totalement. On préfère actuellement le terme web de données car la « sémantique » est une discipline très liée au langage qui implique de plus larges concepts : il s'agit de l'interprétation que l'on donne aux structures syntaxiques (langages de programmation ou structures de données symboliques)[84]. C'est une extension du web actuel, qui ne constitue pas un nouveau web à proprement parler.

Le principe majeur du web de données est le principe AAA[85], qui est de permettre à n'importe qui de dire n'importe quoi à propos de n'importe quoi, et ce « *dans un format lisible pour des machines de manière globalement non ambiguë*. »[86] Pour cela, tout doit être identifié, pas seulement les ressources.[87] Car on ne peut se référer à quelque chose s'il n'existe pas. Or, rappelons-le, l'identification conditionne l'existence d'un objet sur le web. Les résultats de la recherche sur l'intelligence artificielle depuis les années 1950 sont ainsi exploités avec le web de données : il s'agit de permettre à une machine de comprendre le contexte d'un concept en retirant du sens à une déclaration, en la segmentant, et ainsi lui donnant la possibilité de suggérer, recommander des entités connexes et rechercher « intelligemment » des liens de cause à effet. Ça permet d'élargir la capacité des machines à trouver l'information demandée.

Une cellule de recherche se crée en 2006 au sein du W3C nommée SWEO, (pour *Semantic Web Education and Outreach Interest Group)*. Ses objectifs sont de rendre le web sémantique générique, c'est-à-dire présentant les caractéristiques suivantes : pouvoir contenir n'importe quel type de données ; permettre à tout le monde d'y publier ; pouvoir représenter des divergences et informations contradictoires à propos de quelque chose ; constituer un immense graphe global de données ; et s'affranchir de contraintes de vocabulaire, de langue pour représenter les données.[88] D'où l'importance de la construction de passerelles entre les documents et les concepts. L'idée est de permettre une interopérabilité globale et éviter la duplication de l'effort de catalogage et d'attribution de métadonnées. C'est une sorte de méthode de dédoublonnage du savoir mondial. Mais pour cela, il faut

---

[83] ERPA seminar. *Op. Cit.*

[84] HYVONEN, Eero. *Op. Cit.*

[85] Principe AAA : *Anyone can say Anything about Anything*.

[86] BOOTH, David. *Op. Cit.*

[87] AYERS, Danny, VÖLKEL, Max. *Op. Cit.*

[88] BIZER, Christian, HEATH, Tom. *Le web de données*. Pearson France, 2012. [en ligne] Disponible sur https://www.pearson.fr/resources/titles/27440100179400/extras/4746_chap03.pdf [consulté le 09/05/2017]

modeler le monde avec un modèle de données partagées, et créer une infrastructure où les données et les schémas peuvent être publiés, retrouvés et utilisés. Ce que l'on cherche à créer, c'est un « Graphe Global Géant » de ressources connectées.[89]

Principes s'appliquant au web sémantique

La notion de déclaration est primordiale dans le web sémantique. Toutes les informations exprimées doivent être traduites et construites sous forme de déclaration. La phrase de base à laquelle toutes les déclarations peuvent se réduire se décompose en trois segments (un triplet*) qui apportent chacun du sens et induisent un contexte : un sujet, un prédicat, un objet. C'est la base du fonctionnement du RDF (*Resource Description Framework*), le langage du web sémantique. Celui-ci correspond à un modèle de graphe (à différencier d'autres types de modèles de données que nous détaillerons dans la section suivante), c'est-à-dire des connexions entre différents nœuds qui peuvent se faire bi-directionnellement et être tous liés entre eux, à la manière des techniques de cartes conceptuelles de *mind mapping*.

La *Closed World Assumption* (CWA), qui consiste en la présomption que tout ce qui n'est pas connu par le système est par définition faux (préconisée en informatique et notamment dans les bases de données) se doit de devenir une *Open World Assumption* (OWA). Sur le web de données, tout est matière à apprendre, et on ne considère jamais le système comme exhaustif. Cela permet une souplesse et une évolutivité qui a fait partie du succès du web à l'époque de son lancement.

Eero Hyvonen, directeur du Centre d'Humanité Numériques d'Helsinki, identifie plusieurs niveaux de description sur le web de données, en partant de bas en haut[90] :

- Le monde réel : des personnes, des idées, des endroits, des objets, un toucan, une planète…[91]
- La donnée : des documents, des images, des identifiants, des statistiques…
- La métadonnée : des données sur les données tels qu'une notice d'objet, un titre, un lien, une référence…
- L'ontologie : des classes et propriétés utilisées pour un domaine en particulier, un vocabulaire de référence
- La méta-ontologie : l'ontologie globale qui permet de croiser des jeux de données, RDF, OWL…

Cette construction en pyramide montre bien comment le web sémantique, lui-même bâti sur une structure du web articulée en niveaux, se construit à l'aide de briques conceptuelles qui se soutiennent les unes les autres. La liaison des entités et des ressources ne peut se faire que sur une structuration décomposée en strates qui apportent une solidité dans la sémantique théorique et dans la concrétisation pratique et technique du projet.

[89] HYVONEN, Eero. *Op. Cit.*

[90] *Ibid.*

[91] Les toucans et les planètes semblent être des exemples régulièrement cités dans la documentation professionnelle à ce sujet…

Les données liées et l'open data

> « Le terme de données liées doit être compris comme un ensemble de bonnes pratiques pour la publication de données structurées sur le web »[92]

Tim Berners-Lee considère que « *les données liées sont du web sémantique fait correctement*. »[93] Les quatre principes de leur conception sont les suivants[94] :

- L'utilisation d'URI
- L'utilisation d'URI http pour le déréférencement
- La mise à disposition d'informations utiles lors du déréférencement
- La mise à disposition de liens vers d'autres ressources pour permettre l'enrichissement des recherches

Ces données peuvent tout à fait rester internes à une organisation, c'est pourquoi on associe souvent le principe des données liées au mouvement de *l'open-data*, afin que le travail des uns profite aux autres, toujours dans l'objectif de ce graphe mondial idéal. La concaténation des deux, le *Linking Open Data* (LOD) est un projet qui nait en 2007, un an après la constitution du SWEO.

> « L'objectif […] était de commencer le web des données par une identification des jeux de données existants et accessibles sous des licences ouvertes, de les convertir en RDF conformément aux principes des données liées et de les publier sur le web. »[95]

L'idée est d'en faire un réseau hypertextuel exploitable et ouvert, permettant la circulation de l'information selon une structuration appréhendable automatiquement. Le principe des données liées, nous l'avons vu, n'est cependant pas nouveau. Les bibliographies et index de citation sont des formes anciennes de données liées qui trouvent ici avec le paradigme numérique une utilité nouvelle et surtout une concrétisation. Les bibliothèques ne sont d'ailleurs pas en reste sur ce type de mouvement, notamment par la création de groupes de travail tels que le LLD XG (*Library Linked Data Incubator Group*) ou encore la *Bibliographic Framework Initiative* (BIBFRAME) visant à développer le web de données en bibliothèque[96]. Les données liées font ressortir des nouvelles problématiques pour les institutions culturelles: les démarches SEO (*Search-Engine Optimization*) visant à permettre le moissonnage des données dans les sites par les moteurs de recherche (et ainsi conditionnant leur utilisation et leur succès) se développent et favorisent les liens entre les contenus afin d'accorder à certaines ressources un contenu « pivot ». Elles ne sont possibles que lorsque les liens sont stables et pérennes (permettant leur citation sur d'autres pages sans craindre l'erreur 404), et lorsque les ressources sont elles-mêmes déjà pointées par des liens. Ainsi, les anciens catalogues appartenant au web profond deviennent totalement inutilisables, et les bibliothèques se doivent de réutiliser des données existantes ou de mettre en place des méthodes d'interopérabilité.

---

[92] VAN HOOLAND, Seth. VERBORGH, Ruben. *Op. Cit.*

[93] *Ibid.*

[94] BERNERS-LEE, Tim. *Linked Data*. Site W3.org, 2009. [en ligne] Disponible sur https://www.w3.org/DesignIssues/LinkedData.html [consulté le 16/07/2017]

[95] BIZER, Christian, HEATH, Tom. *Op. Cit.*

[96] Voir https://www.w3.org/2005/Incubator/lld/wiki/Main_Page pour le LLD XG et https://www.loc.gov/bibframe/ pour BIBFRAME [consulté le 11/08/2017]

Il est possible de faire de *l'open data* sans faire de web de données, mais la mise en place de données liées permet de développer l'utilisation qui en est faite : ce sont des aspects qui se complètent. Inversement, l'ouverture du *Linking Open Data* est clairement un facteur clé de son expansion. Chaque jeu de données qui s'y ajoute agrandit l'étendue et la complétude du web de données. Nous allons maintenant voir comment sont structurées ces données pour permettre de tels accomplissements.

1.2.4. Structuration de données et modèles de données

L'organisation de la pensée et de la donnée

Il y a plusieurs types de modèles de données, qui sont apparus successivement[97] :

Le modèle de données tabulaire, qui est facile à implémenter et peu couteux (il suffit d'avoir Excel…), est résilient au changement car on peut ajouter et enlever facilement des colonnes. Les formats ouverts utilisés en données tabulaires sont en général CSV (*Comma Separated Values)* et TSV (*Tab Separated Values*). Cependant, ces modèles sont inconsistants, il est difficile de tout harmoniser entre des variations de terminologie, et leur utilisation laisse place à beaucoup de confusions possibles.

Les données tabulaires évoluent dans les années 1970 vers les modèles relationnels développés pour pallier à leur inconsistance et à leur redondance. Avec ces systèmes, on passe à des fichiers binaires propriétaires qui sont très rigides et complexes à mettre à jour, mais qui apportent une efficacité et une rapidité de traitement intéressantes. La gestion amène des liens entre les ressources qui peuvent être nettement plus nombreux que ceux réalisés en deux dimensions dans les données tabulaires. L'interopérabilité n'est cependant pas garantie. Les outils utilisés sont des RDMS (*Relational Database Management Software*) tels que MySQL et Oracle. Ce modèle de données est encore très utilisé actuellement sur le web et dans les organisations, malgré l'incompatibilité avec le web de données.

Les *Meta Markup Language*, ou langages de balisage apparaissent dans la programmation informatique. Il s'agit de délimiter et segmenter des zones de texte pour en structurer le sens pour l'appréhension d'une machine, comme des annotations. XML en est un exemple marquant, tout comme JSON. Ils sont basés sur une organisation hiérarchique, sous forme d'arborescence. Telle information est contenue dans telle balise, elle-même contenue dans telle balise. L'idée est d'obtenir une flexibilité, même au détriment de la rapidité. HTML est d'ailleurs également l'un de ces langages, même si en ce qui le concerne, sa fonction n'est que de représentation, d'esthétique. Il reste néanmoins de la sémantique dans HTML, qui est un mélange de tout cela : c'est sa force et sa limite. Mais ici on perd la fonction de structuration purement sémantique qui était la raison d'être de ces langages, et qui revient à présent sous la forme du web de données, impulsée par le W3C qui était à l'origine même du web. Comme si Tim Berners-Lee était un étudiant qui, quinze jours après avoir rendu son mémoire, se rendait compte qu'il manquait des pages à la version qu'il avait envoyée et essayait de se rattraper…

Les données liées, selon le modèle de graphe nodal, sont un peu organisées comme des bases de données relationnelles, sauf qu'il n'y a pas a priori de structure,

[97] VAN HOOLAND, Seth. VERBORGH, Ruben. *Op. Cit.*

ou que celle-ci n'est pas contraignante. Chaque entité identifiée par une URI peut être utilisée dans plusieurs triplets. RDF est une structuration qui s'intéresse à la logique des données, elle ne correspond pas à l'encodage d'une structure. RDF et XML sont tous les deux des modèles de structuration de l'information, XML est un modèle en arborescence et RDF en graphe. Cependant, XML est un dispositif technique tandis que RDF est un dispositif purement conceptuel (il est d'ailleurs possible d'exprimer RDF en XML).

Il y a également une différence conceptuelle entre les identifiants traditionnels tels que les clés primaires dans une base de données relationnelle, et les identifiants de ressources que sont les URL. Les premiers existent indépendamment des applications web et ne changent pas, tandis que les seconds ont une souplesse de variation et peuvent être reliés à des concepts (par exemple, un article peut être affiché temporairement sur une page d'actualités tout en ayant son identifiant propre).[98]

Développement des modèles adaptés

Les institutions qui sont désireuses de produire des données liées et de les mettre en ligne sont confrontées au problème de la structuration de celles-ci. En effet, l'utilisation d'un modèle spécifique personnalisé est un idéal très gourmand en temps, en moyens humains et financiers, et l'utilisation d'un modèle standard générique est moins chère, mais beaucoup plus compliquée à mettre en place. On peut également combiner plusieurs ontologies pour créer son propre modèle de données. La mise en place du schéma adapté aux besoins de la structure doit se faire dans l'optique d'une interopérabilité maximale avec ceux des autres : même si les gens utilisent le même outil et le même standard, l'interopérabilité n'est pas garantie car ils peuvent très bien l'implémenter de différentes façons selon leurs collections.[99]

Les outils utilisés sont aussi source de questionnements. SPARQL (*Sparql Protocol and RDF Query Language*) est un langage de requêtes dédié aux ensembles de données structurées en RDF. Il est l'équivalent de SQL (*Structured Query Language*) pour les bases de données relationnelles, il permet d'explorer le contenu d'un graphe RDF. Il est quasiment incontournable pour l'utilisation des données liées. Il faut pouvoir mettre en place des descriptions des ressources : les métadonnées sont extensibles à l'infini, chaque représentation peut être documentée par une autre ressource, qui peut également avoir une représentation propre.[100] La solution d'un système de données liées fonctionnel vient souvent d'une part des identifiants adéquats et d'autre part de la qualité des métadonnées.

En outre, Tim Berners-Lee a développé dans son article *Linked Data* un système de notation attribuant des étoiles à la manière des chefs cuisiniers, pour analyser la qualité des données publiées par les organisations[101]:

- 1ère étoile : les données sont disponibles sur le web et en *open-access*,
- 2ème étoile : les données sont disponibles et structurées,

[98] *Ibid.*

[99] BERMÈS, Emmanuelle, ISAAC, Antoine, POUPEAU, Gautier. *Op. Cit.*

[100] VAN HOOLAND, Seth. VERBORGH, Ruben. *Op. Cit.*

[101] BERNERS-LEE, Tim. *Linked Data*. W3.org, 2009. [en ligne] Disponible sur https://www.w3.org/DesignIssues/LinkedData.html [consulté le 11/08/2017]

- 3ème étoile : les données sont disponibles dans des formats open-source plutôt que dans des formats propriétaires,
- 4ème étoile : les identifiants utilisés sont des URI http et leur résolution délivre de ressource en RDF,
- 5ème étoile : les données sont liées entre elles en internet et en externe pour apporter du contexte,

Nous voyons bien qu'ici l'objectif final est l'interopérabilité globale, qu'elle soit le résultat d'un alignement des différentes méthodes afin qu'elles soient compatibles entre elles, ou d'une homogénéisation directement à la création (réutilisation de modèles existants, adaptation).

Conclusion de la partie 1.2

Le web de données, les données liées et l'*open-data* sont donc des concepts qui tendent vers un objectif de gestion globale du savoir : un savoir accessible à tous, partout, tout le temps, en constant accroissement, qui participerait à une omniscience utopique du web. Ils s'approprient et portent une partie des ambitions propres aux institutions culturelles, bibliothèques, archives, musées, et sont sujets à un intérêt grandissant. La modélisation des données est quant à elle le cœur du processus, qui conditionne la réussite entière du projet. Avec RDF, il n'y a pas de limites aux connexions possibles. Il accepte tous les points de vue, et n'apporte pas de test de véracité ou de qualité de l'information. En conséquence, certains triplets peuvent se contredire ou ne rien vouloir dire du tout. L'assurance de qualité et de véracité est forcément apportée par un contrôle humain, et cela est assez difficile à mettre en place. Les problèmes liés à la véracité d'un triplet ne peuvent être dénoués qu'avec une vision du contexte.[102] Nous développerons ces aspects plus en détail dans la partie 3.

[102] WILLER, Mirna, DUNSIRE, Gordon. *Op. Cit.*

1.3. IDENTIFIER NUMERIQUEMENT, DANS QUEL CONTEXTE ET POUR QUEL OBJECTIF ?

1.3.1. L'architecture REST et les API

L'architecture idéale du web

Le *Representational State Transfer*, ou REST, est un style d'architecture pour les systèmes d'information. Il est créé en 2000 par Roy Fielding et comporte des prérequis applicables sur le web. Fielding était assez connu déjà de par son implication dans le développement du web, notamment http. On nomme RESTful les systèmes qui fonctionnent selon ce type d'architecture. Le principe, dans un contexte web, est d'utiliser systématique le protocole http ainsi que les URI comme identifiants, ce qui permet d'avoir « *un système universel d'identification des éléments de l'application* »[103]. Chaque ressource est identifiée de manière unique. La transparence des processus et la simplicité de leur fonctionnement est le maître-mot du REST : si une représentation est encodée en UTF-8, cela doit être explicite. Les ressources ont des représentations bien définies, et leur manipulation se fait via celles-ci.[104]

L'implémentation* du REST est en outre facile à mettre en place : il y a un très faible coût pour l'application, puisque l'on ne va pas « développer » spécifiquement une application pour ce service ; la maintenance en est facilitée ; côté client le développement et la maintenance sont faibles également, ce qui de manière globale garantit une facilité d'adoption très intéressante pour les structures culturelles qui souhaitent publier leurs jeux de données en ligne.

Une architecture en REST permet d'accéder à un document dans toutes ses représentations via un seul identifiant majeur identifiant la ressource en tant que concept : c'est ce qu'on appelle la négociation de contenu*.

> « Beaucoup de gens ne savent pas que les URL n'identifient pas des fichiers mais des ressources. Cette conception amène la croyance erronée que l'accès html est forcément différent de l'accès aux versions en JSON et en RDF. »[105]

Concrètement, cela oblige à attribuer plusieurs URI : une pour la ressource, qui est générale et qui identifie le concept, l'entité, et une pour chacune des représentations de cette entité (RDF, JSON, HTML, etc.). Cela nécessite un développement « par le bas » (*bottom-up*) : partir de l'application finale désirée et du besoin pour créer un « hub » de métadonnées qui va permettre le lien entre les différentes URI créées.[106]

La négociation de contenu

Certains navigateurs gèrent le RDF et le html et indiquent leurs préférences dans leur header, plus spécifiquement dans la partie « accept » (par exemple Mozilla

---

[103] *Representational State Transfer*. Site Wikipédia [en ligne] Disponible sur https://fr.wikipedia.org/wiki/Representational_state_transfer [consulté le 17/07/2017]

[104] *Ibid.*

[105] VAN HOOLAND, Seth. VERBORGH, Ruben. *Op. Cit.*

[106] BERMÈS, Emmanuelle, ISAAC, Antoine, POUPEAU, Gautier. *Op. Cit.*

avec son extension Tabulator)[107]. Ainsi, la négociation de contenu fonctionne selon les préférences indiquées dans les requêtes: le serveur de contenu va récupérer cette information lors de la première requête afin de pouvoir s'adapter au contexte et, grâce à cette information, fournir une représentation en HTML si c'est un utilisateur qui le demande via son navigateur, ou fournir un flux encodé suivant une des sérialisation de RDF si c'est une machine qui le demande en spécifiant explicitement ce format.[108]

La préférence pour un format ou pour un autre est indiquée par une valeur, « q », plus elle est élevée plus elle influe sur le choix.[109] Le choix est donc plus « fin », si la page HTML est simplement une copie bas de gamme alors le serveur renverra le RDF (sauf en cas de préférence explicite du client) ; inversement, si le RDF n'est pas complet ou vient directement du HTML, alors le serveur renverra la représentation en HTML (sauf, là encore, en cas de préférence du client).[110]

Une fois la requête reçue, le serveur redirige le client vers une URL où se trouve la représentation correcte. Cette redirection a pour nom de statut code *302 Found*. Le client envoie alors une nouvelle requête http au serveur de la nouvelle URL qui donne cette fois accès à la ressource dans la représentation souhaitée.[111]

Les API

Les *Application Programming Interface* (API) ou *Interfaces de programmation applicative* sont des interfaces à travers lesquelles un fournisseur de contenu offre un service à d'autres applications/logiciels. Le principe des API dans notre cas particulier est donc de déployer une interface entre le navigateur de l'utilisateur et la ressource, et ce, afin de récupérer les requêtes et redistribuer lui-même en fonction de la demande la version RDF, HTML, ou JSON d'un document. « *Une API branchée sur une base de données fonctionne donc de la manière suivante : vous lui fournissez une URL contenant un ou plusieurs paramètres ; elle vous renvoie un fichier structuré (XML ou JSON) avec une ou plusieurs réponses.* »[112]

Dans le cas d'une architecture REST, l'API utilisée pour toutes les représentations dédiée à l'appréhension humaine est toujours le navigateur. D'autres fournisseurs de contenus préfèrent développer eux-mêmes une API qui va « aiguiller » les utilisations faites et restreindre en termes d'accès à certaines ressources, parfois en proposant à l'utilisateur de s'identifier avec un compte et un mot de passe[113].

---

[107] AYERS, Danny, VÖLKEL, Max. *Op. Cit.*

[108] BERMÈS, Emmanuelle, ISAAC, Antoine, POUPEAU, Gautier. *Op. Cit.*

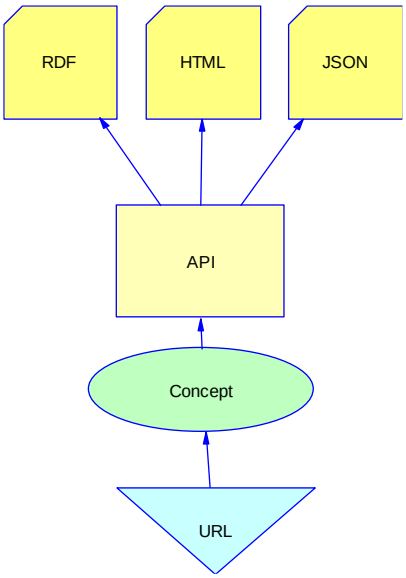
[109] Par exemple, une préférence pour le html pourra être exprimée telle que html q=1.0 et RDF q=0.7.

[110] AYERS, Danny, VÖLKEL, Max. *Op. Cit.*

[111] Ministère de la Culture et de la Communication. *Op. Cit.*

[112] CAVALIE, Etienne. *Utiliser les SPARQL Endpoint comme si c'était des API*. Site Bibliotheques.wordpress.com, 2012. [En ligne] Disponible sur https://bibliotheques.wordpress.com/2012/11/22/utiliser-les-sparql-endpoint-comme-si-cetait-des-api/ [consulté le 17/07/2017]

[113] Ceci dit, c'est également faisable en REST en limitant l'accès à des utilisateurs authentifiés via une clé.



C'est le cas notamment de la *Digital Public Library of Australia* (DPLA) qui n'est pas structurée en REST, et qui propose en outre une partie de ses ressources en libre accès en HTML. Si l'on souhaite accéder à la version JSON de celles-ci, il faut s'enregistrer pour accéder à l'API DPLA spécifique. Cela demande en sus une procédure longue et fastidieuse nécessitant de trouver la syntaxe exacte des demandes en JSON dans leur documentation, puisque c'est une API qui leur est propre, et ensuite trouver l'URL spécifique de l'objet que l'on recherche en particulier à partir de l'URL en HTML, etc. Tout cela doit se faire manuellement et individuellement pour chacune des ressources, ce qui limite le moissonnage automatique et la réutilisation des données. De plus, on voit bien que connaître le fonctionnement de l'API dédiée d'une certaine base de données, (ici, celle de la DPLA) ne donne pas un savoir universel sur toutes les autres bases de données que l'on pourrait utiliser. Chacune est façonnée selon son propre modèle et cela participe au cloisonnement des informations.[114]

Europeana fait également partie de ces fournisseurs de contenu qui fonctionnent avec une API personnalisée, mais celle-ci est en REST et Open-Source. Elle vient d'ailleurs tout juste d'être récompensée du prix de l'API gagnante dans la catégorie des API de données, du prestigieux évènement API World 2017 ayant lieu dans la Silicon Valley en Calfiornie[115]. Le fondateur de l'évènement, Geoff Domoracki, déclare que « *L'API REST d'Europeana est un merveilleux exemple montrant comment l'industrie technologique réalise de plus en plus que les API ne sont pas juste une manière d'intégrer des outils et des applications, mais qu'elles sont la nouvelle façon de faire du commerce et de créer des nouvelles technologies révolutionnaires.* »[116]

Que choisir ?

Les systèmes qui ne sont pas construit en REST font face à certaines difficultés, notamment le fait que les identifiants de chaque représentation d'une même ressource ne soient pas liés à un identifiant « maître », rendant l'interopérabilité difficile. Les URI identifient des représentations plutôt que des ressources, nous avons donc besoin d'être vigilants sur la construction de celles-ci. En outre, les URL JSON incluent souvent des informations sensibles touchant au cœur du fonctionnement du système, conséquence de l'utilisation des URL non pas comme identifiants pour des ressources mais comme des instructions pour programme. On ne peut donc pas forcément les diffuser tels qu'elles. Inversement, les RESTful API ont des détracteurs de leur côté, qui estiment qu'elles ne sont pas la panacée en termes de fonctionnement technique (difficulté à « débuguer », le

[114] VAN HOOLAND, Seth. VERBORGH, Ruben. *Op. Cit.*

[115] *Europeana REST API wins the 2017 Data APIs category award*. Pro.europeana.eu [en ligne] Disponible sur http://pro.europeana.eu/pressrelease/europeana-rest-api-wins-the-2017-data-apis-category-award [consulté le 16/08/2017]

[116] *Ibid.*

vocabulaire REST n'est pas assez riche pour une API, une confusion terminologique sur ce que veut vraiment dire une API en REST, etc.).[117]

D'autres types d'architectures permettent de gérer les données : l'architecture SOA (*Service Oriented Architecture* ou Architecture orientée services) est une architecture dont le but est d'établir des connexions entre les bases de données une à une. Cela repose sur l'alignement des processus et non des données. Cependant cela donne parfois lieu à des architectures informatiques complexes, difficiles à maintenir et peu évolutives.[118] SOAP (*Service Oriented Architecture Protocol*) est à l'architecture SOA ce qu'http est à l'architecture REST : un protocole servant au transport et à la structuration des données. Il y a en outre d'autres types d'architectures plus spécifiques, notamment l'architecture *Business Intelligence*, intéressante dans le cas de données quantitatives pour les statistiques, les vues de données… au sein des entreprises notamment ; ou encore le *Master Data Management* proposant une base pivot entre toutes les données.[119] Cette diversité permet aux organisations d'adapter leurs systèmes à leurs besoins précis, même si les avantages et les inconvénients doivent être étudiés avec attention lors de l'implémentation et ne doivent pas entrer en contradiction avec les objectifs initiaux.

1.3.2. Vocabulaires contrôlés et référentiels

Structurer le monde, une volonté ancienne

L'ontologie* est une branche de la philosophie qui s'attache à l'étude de l'existence, de la structure et de la nature des choses telles qu'elles sont. Aristote lui-même s'intéressait à ce domaine dans sa catégorisation du monde, la catégorie étant étymologiquement « *le mode d'accusation de l'être* », qui lui permet de diviser en 10 les parties constituantes de la réalité et des concepts. En effet, nous l'avons vu, structurer le monde, la connaissance et apporter un cadre de segmentation par typologies/ontologies est un moyen de faciliter son accès et son appréhension. Le savoir se doit d'être transmis non pas comme un tout mais par « paquets » bien déterminés, reliés les uns aux autres.

Les bibliothèques s'en chargent par la création de référentiels*, les archives par celle des séries de « domaines », les musées par des collections (correspondant souvent aux classifications au sein des savoirs notamment la biologie et le vivant qui en sont des exemples criants). Chacun a son organisation propre et comporte ses propres clés de compréhension du monde, son propre point de vue. Les points d'entrée peuvent ainsi être divers : thématiques, chronologiques, topographiques, subjectifs, scientifiques, alphabétique, etc. La notion d'ordre et de « cases » est consubstantielle à la notion de pensée. Le cerveau humain ne peut concevoir un concept en globalité sans en percevoir une certaine structure, une certaine organisation interne.

Ainsi, les besoins de description des entités sont également tout à fait différents : si en bibliothèque on privilégiera les titres et auteurs pour désigner un ouvrage, par exemple, en archives ce sont les contextes qui seront utilisés, et en

[117] MIKOWSKI, Michael S. *RESTful APIs, the big lie, Why you might benefit from letting this popular paradigm rest in peace*. Blog mmikowski.github.io, 2015. [en ligne] Disponible sur https://mmikowski.github.io/the_lie/ [consulté le 17/07/2017]

[118] BERMÈS, Emmanuelle, ISAAC, Antoine, POUPEAU, Gautier. *Op. Cit.*

[119] *Ibid.*

musée et dans d'autres secteurs il y aura encore d'autres objectifs et modes de fonctionnement. Sur le web de données, l'enjeu est le même.

Différencier les différentes terminologies employées

Nous retrouvons plusieurs terminologies qui distinguent des méthodes d'organisations plus ou moins différentes : les vocabulaires contrôlés, les ontologies, et les référentiels. Ce sont des notions très importantes dans l'objet qui nous concerne, et sur lesquelles il est utile de s'accorder.

Antoine Isaac, manager en recherche et développement à Europeana (et participant au LLD XG), considère qu'il peut y avoir trois types de référentiels dans la gestion de l'information [120]:

- Premièrement, cela peut être des éléments de métadonnées constituées de classes et propriétés pour les descriptions que l'on va nommer ontologies. Celles-ci fournissent des règles de raisonnement, des axiomes. « *Les ontologies pour le web de données introduisent et définissent de façon formelle (utilisant les langages RDFS et OWL) les éléments nécessaires à l'expression de (méta)données.* [121]»
- Deuxièmement, cela peut être des vocabulaires de valeurs appelés aussi vocabulaires contrôlés ou systèmes d'organisation des connaissances (KOS, *Knowledge Organisation System*), tels que les thésaurus, les fichiers d'autorité, ou encore des bases de connaissances pouvant être volumineuses (VIAF, Dewey, Rameau, GeoNames…). Ce ne sont pas des modèles de données, ils ne les structurent pas, mais ils proposent en revanche des occurrences validées et cohérentes les unes entre les autres.
- Enfin, ils peuvent prendre la forme de jeux de données de très bonne qualité, qui le seraient assez pour devenir des « références ». Tout jeu peut être réutilisé, mais certains font néanmoins de meilleurs candidats (par exemple, DBpedia, Freebase, etc.).

Pour résumer, si les ontologies régentent la structure même des données en créant des champs, les KOS eux, viennent les remplir.

Outils utilisés et exemples

Parmi les outils liés aux ontologies formalisées, on retrouvera : OWL (*Web Ontology Language*) un langage d'ontologie RDF qui permet la déclaration d'équivalence entre deux notions, la hiérarchisation et la distinction entre deux classes ; RDFS, très rigide en terme d'expressivité mais qui permet également de hiérarchiser les classes et propriétés de RDF ; et enfin SKOS (*Simple Knowledge Organization System*) une recommandation du W3C censée faciliter l'échange de données sémantiques en proposant des équivalences conceptuelles, notamment avec OWL, et qui permet d'utiliser des thésaurus via un format RDF.

Les vocabulaires contrôlés, ou KOS, sont composés d'un corpus de termes qui sont eux-mêmes des identifiants, puisqu'ils sont uniques dans leur domaine et dans leur type. Une fois exprimés en URI, ils deviennent juste des identifiants de données

[120] ISAAC, Antoine. *Les référentiels: typologie et interopérabilité*. Séminaire IST Inria : le document numérique à l'heure du web de données, Carnac 2012. [en ligne] Disponible sur https://hal.inria.fr/hal-00740282v1 [consulté le 09/05/2017]

[121] *Ibid.*

liées, que l'on peut utiliser tels quels moyennant une complétude dans la syntaxe.[122] Ils prennent de l'intérêt et de la valeur à force d'être réutilisés. Malheureusement, ce n'est pas vraiment le cas actuellement :

> « Les vocabulaires contrôlés sont comme des sous-vêtements. Tout le monde pense que c'est une bonne idée mais personne ne veut utiliser ceux des autres. »[123]

Le problème des vocabulaires contrôlés, c'est qu'il est difficile justement d'ajuster le niveau de contrôle. Certains peuvent clairement être trop ambitieux et rigidifier les possibilités d'utilisation mais surtout d'évolution. Ils jouent néanmoins un rôle important dans une partie des principes du web sémantique de Tim Berners-Lee : ils permettent de procurer des informations connexes à un usager qui consulte une ressource ou un concept, et il inclut d'autres liens vers d'autres URI de manière à ce qu'il puisse découvrir d'autres ressources.[124] Il serait donc intéressant que ces réutilisations se développent.

L'alignement et l'identification en vue de l'interopérabilité

Cliff Morgan et Norman Paskin (notamment un des fondateurs du système DOI), quant à eux, voient les dictionnaires de données et les tables de cartographies de métadonnées comme essentielles à l'interopérabilité des différents domaines. Mais cela nécessite forcément un alignement des référentiels. Cette étape, primordiale, est très compliquée à gérer. En effet, en ce qui concerne les KOS, souvent les contenus sont trop volumineux, moins bien structurés que les ontologies et ils sont constitués de libellés très hétérogènes. L'alignement multilingue est notamment un gros problème, la langue étant ce qu'elle est : ambigüe, évolutive, insaisissable, redondante.[125]

De même, dans les ontologies, si la réutilisation n'est pas possible à la création du système, un alignement peut être réalisé à postériori. Le principe sera alors de faire des déclarations d'équivalence d'une classe à une autre afin de recouper les données et créer des passerelles. L'alignement de référentiel en général peut se faire par deux moyens différents : par « *hub and spoke* », méthode consistant à mettre en place un référentiel central qui fait le lien avec tous les autres (par exemple, le fichier virtuel d'autorité du VIAF) ; ou alors par « structures de paires », qui est plus complexe à mettre en place et dont le principe est de relier chaque référentiel à tous ses voisins. La première méthode est plus facile et plus souple, mais cependant plus risquée car il faudra rentrer dans une granularité de spécialité parfois importante.

En soi, l'objectif global est donc d'identifier au mieux les entités et les référentiels pour pouvoir permettre à d'autres de les réutiliser, de faire des liens avec eux et de publier leurs données sur le web de manière correcte. Les identifiants ont un rôle crucial à jouer, même au niveau structurel. Le conseil technologique du secteur public du Royaume-Uni préconise d'ailleurs à ses agents l'emploi de 6 types d'URI par niveau conceptuel, correspondant à 5 types de ressources [126]:

---

[122] PASKIN, Norman, RUST, Godfrey. *Op. Cit.*

[123] VAN HOOLAND, Seth. VERBORGH, Ruben. *Op. Cit.*

[124] *Ibid.*

[125] ISAAC, Antoine. *Op. Cit.*

[126] DAVIDSON, Paul, CIO Sedgemoor, District Council. *Designing URI Sets for the UK Public Sector*. Chief Technology Officer Council, 2009. [en ligne] Disponible sur

- Les URI « identifiant » comprenant les objets physiques et abstraits du monde réel,
- Les URI « document » comprenant les *web documents* en tant qu'entité,
- Les URI « représentation » comprenant les occurrences de chaque format de fichier par entité,
- Les URI « liste » correspondant à l'identifiant de l'index de toutes les URI présente dans l'ensemble,
- Les URI « ontologie » qui identifient les éléments des ontologies, incluant les relations entre les choses identifiées.

Nous voyons bien ici l'importance que peuvent prendre les identifiants dans le graphe global des données liées, s'ils sont la seule vraie voie d'accès aux entités mais également aux éléments qui les structurent.

1.3.3. Quelques exemples de systèmes et leurs outils

Europeana, la riposte européenne

Initiée en 2005 et ouverte en novembre 2008, Europeana est un projet très ambitieux de la Commission européenne sur un budget total d'environ 100 milliards d'euros. L'objectif était de mettre à disposition en données liées le maximum de ressources patrimoniales, artistiques, scientifiques, muséales et historiques de l'Europe sur une base de données à grande échelle, de très nombreuses institutions culturelles étant invitées à se joindre au projet et à donner accès à leurs données. D'autre part, Europeana se plaçait à l'époque en travers de la route du géant Google, parti sur une logique d'expansion semblant irrépressible, et constituait une riposte Européenne. Elle poursuit ainsi un objectif d'open-data assumé, notamment avec l'adoption d'un *Data Exchange Agreement* (DEA) avec les fournisseurs de données, visant à lui permettre de mettre à disposition les métadonnées des collections dans une licence *Creative Commons*, les passant dans le domaine public[127]. Actuellement, elle culmine à plus de 53 millions de documents numérisés mis en ligne, provenant de 3300 institutions de tous types sur une couverture temporelle très large (depuis l'Antiquité à nos jours). « *C'est actuellement le plus grand, sinon l'unique projet culturel à l'échelle du continent.* »[128]

Le modèle de données d'Europeana (l'EDM, ou *Europeana Data Model*) fait donc cohabiter des données provenant de sources très diverses : à la fois des bibliothèques, des archives, des musées, etc. Les équipes en charge du projet réutilisent les métadonnées existantes sur une ontologie de haut niveau et incluent plusieurs ontologies, SKOS pour les concepts, FOAF pour les personnes, et Dublin Core pour les métadonnées descriptives de base.[129] Vingt-et-une langues coexistent sur cette bibliothèque numérique, qui valorise la réutilisation en proposant des

https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/60975/designing-URI-sets-uk-public-sector.pdf [consulté le 18/05/2017]

[127] *Europeana – A European Digital Library for all.* Ec.europa.eu [en ligne] Disponible sur https://ec.europa.eu/digital-single-market/en/europeana-european-digital-library-all [consulté le 16/08/2017]

[128] *Europeana Collections, Le patrimoine de l'Europe en ligne*. Bnf.fr [en ligne] Disponible sur http://www.bnf.fr/fr/collections_et_services/bibliotheques_numeriques_gallica/a.europeana_bib_num.html [consulté le 16/08/2017]

[129] BERMÈS, Emmanuelle, ISAAC, Antoine, POUPEAU, Gautier. *Op. Cit.*

modes de recherche originaux tels que la recherche dans les images par nuancier de couleur.

Dans leur document qui traite des solutions d'interopérabilité pour les administrations publiques européennes[130], les auteurs de l'ISA (*Interoperability Solutions for European Public Administrations*) parlent du cas Europeana et font à son sujet des remarques intéressantes : elle aurait fait face à des difficultés vis-à-vis de l'attribution de ses identifiants. A la base, ceux-ci sont attribués de manière séquentielle et automatique, incrémentés au moment de l'intégration de nouveaux éléments, sur le schéma d'URI suivant : http://data.europeana.eu/item/identifiant-de-la-collection/identifiant-de-l'item. Cependant, certaines collections ont été réintégrées après coup et donc se sont vues attribuer deux fois des identifiants. « *Il est évident que les systèmes d'attribution automatique d'URI devraient avoir des moyens d'éviter ce genre d'écueil à l'avenir*. »[131] Il faudrait, pour cela, qu'ils eussent dans les métadonnées un champ Identification auquel on peut faire référence lorsque le système secondaire récupère les données. Cependant, il est possible de réaliser a postériori une fusion des ressources dont les données sont identiques pour ne garder que la première URI assignée. C'est par contre sujet à erreur et assez délicat à mettre en place.[132]

En conséquence, même si nous privilégierons d'autres études de cas dans la seconde partie de ce mémoire, Europeana semble être un projet de grande ampleur qu'il aurait aussi été intéressant d'étudier sur leur mise en place d'identifiants, surtout pour les obstacles qu'ils ont pu rencontrer et qui auraient constitué un retour d'expérience.

Discrimination homme/machine : Wikipédia et DBpedia

Wikipédia, que l'on ne présente plus, est un projet fascinant. Fonctionnant sur le principe du wiki (participatif), cette encyclopédie « du peuple » est un lieu ouvert, très riche en informations, et autonome. Wikipédia possède des URI non pérennes assumées, elles sont systématiquement basées sur le titre d'une page et sont créées automatiquement. Elles sont conçues pour « parler » aux humains, pas aux machines, pour lesquelles la version DBpedia existe.

DBpedia est la version « web de données » de Wikipédia. Elle possède une déclinaison française à l'initiative de l'INRIA, Semanticpedia, qui concerne toutes les données en français du site. Elle propose les connaissances de Wikipédia structurées en RDF, que l'on recherche via des requêtes SPARQL.

D'un côté, donc, la souplesse des URI de Wikipédia permet une adaptabilité aux évolutions des concepts, propres à des utilisateurs humains qui ont une grande flexibilité dans leur pratique de la langue et surtout qui savent très bien distinguer grâce au contexte. Néanmoins, afin de désambiguïser les homonymies, des éléments sont présents entre parenthèses à la fin des URI. [133] Par exemple http://fr.wikipedia.org/wiki/Freebase_(web) correspond à la page Wikipédia du

---

[130] ARCHER, Phil, GOEDERTIER, Stijn, LOUTAS, Nikolaos. Interoperability Solutions for European Public Administrations. *D7.1.3 - Study on persistent URIs, with identification of best practices and recommendations on the topic for the MSs and the EC*. ISA, 2012. [en ligne] Disponible sur https://joinup.ec.europa.eu/sites/default/files/D7.1.3%20-%20Study%20on%20persistent%20URIs_0.pdf [consulté le 19/05/2017]

[131] *Ibid.*

[132] *Ibid.*

[133] *Ibid.*

projet de *Linked data* Freebase (clôturé depuis 2015 lorsqu'il a été acquis par Google), tandis que l'URI http://fr.wikipedia.org/wiki/Free_base correspond à la page de description d'un psychotrope illicite. La confusion entre les deux notions aurait été inopportune.

De l'autre côté, sur DBpedia, les URI sont figées. La recherche d'URI se fait via les requêtes, ou via les URI lorsqu'ils sont connus. L'utilisateur humain qui souhaite chercher manuellement peut également « deviner » l'URI d'une ressource en utilisant l'espace de nom http://dbpedia.org/resource/ puis en ajoutant le nom de sa recherche exacte, en remplaçant les espaces par des *underscores* (tirets bas « _ »). Les recherches via des requêtes SPARQL ne sont cependant jamais exhaustives : le moteur s'arrête de chercher après un certain contingent de retours. Il est donc assez difficile de distinguer la base réelle de connaissances qui y est contenue, et cela pose aussi le problème de l'obsolescence ou l'incorrection des données.[134]

Cette frontière discriminant humain/machine tend cependant à se brouiller puisque Wikipédia souhaite tendre vers des URI plus pérenne, afin de passer d'un service anthropo-centré à un service qu'humain et machines peuvent tous les deux utiliser, rendant ainsi critique le management des URI.

Le RDD d'INDECS, une base généraliste pour la gestion des droits

Le projet INDECS « *Rights Data Dictionnary* », ou dictionnaire de données sur les droits est un projet intéressant pour son objectif global : il s'agit, comme son nom l'indique, de la construction et de l'implémentation d'un référentiel de données relatives aux droits globaux de par le monde pouvant être commun à la majorité des entreprises et des situations. C'est un domaine délicat puisque beaucoup de référentiels sont déjà utilisés, souvent des solutions individuelles à certaines organisations. De plus, les droits sont des éléments difficiles à gérer, de par leur complexité, leur évolutivité dans le temps ainsi que leur éventuelle mixité avec d'autres informations.[135]

Imaginé en 2001 par le consortium <indecs> rdd, le projet a été conçu par Rightscom pour que chaque terme référencé ait un identifiant unique et une généalogie, c'est-à-dire une relation avec les autres termes du vocabulaire, sur l'image des données liées. Ce « dictionnaire » est légalement neutre, c'est-à-dire qu'il ne sert pas à définir des termes légaux et donc est utilisable dans tous les pays. Il est aussi neutre sur le plan du domaine d'activité, permettant ainsi son utilisation dans des situations d'échanges commerciaux très divers où la gestion des droits est requise[136].

En outre, c'est une base conceptuelle solide avec un modèle de données réfléchi, il est inclusif (les termes des autres référentiels peuvent être ajoutés), et sa granularité est forte et souple. En termes de métadonnées, il permet la coexistence de plusieurs types de métadonnées. Il ne faut cependant pas le confondre avec un

[134] VAN HOOLAND, Seth. VERBORGH, Ruben. *Op. Cit.*

[135] PASKIN, Norman. *Towards a rights data dictionary, Identifiers and semantic at work on the net*. EPS, 2002. [en ligne] https://www.doi.org/topics/020522IMI.pdf [consulté le 29/11/2016]

[136] AGNEW, Grace. *Digital Rights Management: a librarian's guide to technology and practise*. Elsevier, 2008. 452p.

REL (*Rights Expression Language*), qui lui utiliserai les termes définis dans le Rights Data Dictionary.[137]

Il a depuis été utilisé dans la construction et l'enrichissement d'applications tels que le schéma de métadonnées du système DOI (*Digital Object Identifier*), du DDEX (*Digital Data Exchange* concernant l'industrie de la musique), ou encore le cadre de RDA/ONIX[138] pour la catégorisation des ressources.[139]

Le système de NER

Les outils de NER (*Named-Entity Recognition*) sont des logiciels qui identifient des termes spécifiques dans des textes non structurés et qui les désambiguïsent en les comparant à des bases de connaissances telles que DBpedia.[140] Leur utilisation est connexe aux pratiques de publication de données liées qui participent à leur expansion, c'est pourquoi nous l'évoquons ici. Elles peuvent notamment servir à accélérer l'inclusion des jeux de données individuels à une base plus générale sur le web de données, de par cette identification automatique des concepts et leur liaison.[141]

Le taux de fiabilité d'un système de NER (sa précision et son rappel) peut être mesuré en comparant ses résultats à des résultats obtenus par une reconnaissance humaine de termes. Ceux-ci sont des échantillons élaborés manuellement, nommés *Gold Standard Corpus*, et ils constituent l'idéal que l'on souhaite obtenir de manière automatique.

Un exemple de service de NER est DBpedia Spotlight qui est open source. Seth Van Hooland et ses collaborateurs exhortent fortement dans leur article sur le sujet des NER[142] les institutions culturelles et les organisations à utiliser ce type d'outils pour optimiser leur travail de publication de données, afin de faire face aux géants du domaine (Google, Facebook) qui, pour l'instant, imposent les règles et constituent le web de données actuel.

1.3.4. Les projets de Linked Enterprise Data (LED)

Pourquoi s'y intéresser et quels objectifs ?

Ici, et pour parfaire notre plongée au cœur des notions, nous allons explorer succinctement les projets de données liées, voire de web de données mis en place côté privé, dans les organisations à but lucratif. Cette optique est intéressante dans notre développement car elle montre comment il est possible de tirer profit de ces

---

[137] <indecs> rdd Consortium. *<indecs> rdd White Paper, a standard Rights Data Dictionary*. 2002. [en ligne] Disponible sur http://www.doi.org/topics/indecs-rdd-white-paper-may02.pdf [consulté le 19/07/2017]

[138] RDA/ONIX est « *un cadre commun pour la catégorisation des ressources selon leur contenu et leur présentation matérielle, afin de faciliter les échanges et la réutilisation des descriptions de ressources entre les deux communautés* » RDA et ONIX. BnF. RDA (Ressources : Description et Accès). [en ligne] Disponible sur http://www.bnf.fr/fr/professionnels/rda/s.rda_objectifs.html [consulté le 14/08/2017]

[139] DOI. *Factsheet, The indecs framework*. [en ligne] Disponible sur https://www.doi.org/factsheets/indecs_factsheet.html [consulté le 19/07/2017]

[140] VAN HOOLAND, Seth. VERBORGH, Ruben. *Op. Cit.*

[141] VAN HOOLAND, Seth, DE WILDE, Max, VERBORGH, Ruben, STEINER, Thomas, VAN DE WALLE, Rik. Exploring Entity Recognition and Disambiguation for Cultural Heritage Collections. *Literary and Linguistics Computing*, 2014. 18p. [en ligne] Disponible sur http://freeyourmetadata.org/publications/named-entity-recognition.pdf [consulté le 19/07/2017]

[142] *Ibid.*

technologies et d'une gestion de l'information optimisée non seulement dans un cadre « culturel » et open-source mais également dans le cadre entrepreneurial. Le cloisonnement des domaines met généralement un frein aux bénéfices exploitables d'une technologie quand chacun développe ses projets dans son coin et que les innovations et les bonnes pratiques ne sont pas « transectorielles ».

Le LED correspond en réalité à l'utilisation des technologies du web de données dans le cadre des besoins strictement internes à une entreprise. Les données produites en entreprise sont, par essence, hétérogènes. Elles sont issues d'une prolifération d'application métiers qui bourgeonnent à l'heure où la dématérialisation est bien entamée mais pas toujours selon des logiques *records management*. Auparavant, il était facile de construire un système logiciel pour n'importe quoi, un produit, un besoin, mais maintenant que l'exigence de vitesse de réactivité s'est accrue il devient lourd de produire et de gérer ces mêmes logiciels.[143] Les processus métiers ne sont pas forcément eux-mêmes très efficaces et il peut exister des incohérences et des doublons dans les données. L'objectif du LED est donc de fluidifier l'expérience utilisateur, perturber le moins possible les applications métier et créer de nouveaux services et outils de consultation.[144]

En outre, les valeurs du LED sont davantage axées sur la sécurité des données qu'elles ne le sont sur le web et dans le cas des institutions culturelles, car beaucoup de décisions cruciales sont prises uniquement sur l'information reçue. L'information critique au succès des entreprises est plus difficile à trouver, intégrer et utiliser. Les entreprises doivent plus que jamais développer leur agilité, ce qui implique également une agilité des systèmes. Ici, il est donc question de la survie même de l'entreprise dans son environnement économique et concurrentiel. Les enjeux sont donc équivalents, sinon plus forts, que ceux des institutions culturelles qui gèrent la connaissance.

Organisation des connaissances en entreprise

> « La clé du LED repose dans la capacité à réutiliser les données des différentes applications qui constituent le système d'information, tout en respectant les besoins métier particuliers qui justifient l'existence de bases de données diverses. De la même manière que sur le web de données ouvert, on va construire une interopérabilité basée sur des liens, et non sur l'adoption d'un format unique ou d'un dénominateur commun. »[145]

Il y a une tension fondamentale en entreprise entre la vue globale et les vues locales. La représentation des données change avec son propriétaire, c'est ce qui rend la question de l'interopérabilité si épineuse. Les approches logicielles et organisationnelles abordées pour gérer cette question peuvent être de plusieurs types [146]: les entrepôts de données basés sur des bases de données relationnelles, constituant des réceptacles que l'on peut interroger ; les approches MDM (*Master Data Management*) que nous avions évoquées plus haut concernant l'architecture de données, correspondant à une approche holistique de la donnée d'entreprise ; les réceptacles de métadonnées, utilisées comme des liens entre les différentes sources

[143] ALLEMANG, Dean. Dans WOOD, David. *Linking Enterprise Data*. Springer, 2010. Chapitre "Semantic Web and the Linked Data Enterprise". p3-23.

[144] BERMÈS, Emmanuelle, ISAAC, Antoine, POUPEAU, Gautier. *Op. Cit.*

[145] *Ibid.* p159

[146] ALLEMANG, Dean. *Op. Cit.*

de métadonnées (flexibles, mais souvent construite avec des formats propriétaires) ; les vocabulaires contrôlés que nous avions également vus ; ou encore le *Natural Language Processing*, forme simple d'extraction de concepts. L'étape d'indexation sur laquelle repose tout le système est cependant souvent manuelle, et soumise à la bonne volonté du producteur de la donnée. En effet, l'une des manières de s'assurer d'un traitement optimisé de la donnée d'entreprise serait de garantir que l'action de création soit directement et intimement liée à celle de son partage. « […] *à l'entrée dans la vie active d'une génération qui a grandi avec Wikipédia, de plus en plus de techniciens se rendent compte qu'ils ne peuvent pas faire l'économie de cet effort* [de rendre leurs données accessibles et partageables]. *Une minute d'effort de partager l'information résulte en des heures de gagnées en utilisant l'expertise des collègues* »[147].

Ces propos sont à nuancer cependant à la lumière d'autres arguments, d'une part ceux de Cory Doctorow[148], d'autre part de ceux de Catherine Legg[149]. En effet, l'attribution de métadonnées par son producteur, l'indexation manuelle, le nommage et enfin la publication ne sont peut-être pas la manière la plus sûre d'obtenir des jeux de données propres et interopérables dans tous les cas. Selon Doctorow, les gens peuvent mentir, être feignants ou stupides, ils ne se connaissent pas eux-mêmes et les guides ne sont jamais neutres[150]. Enfin, l'idée que la donnée ne peut pas être mieux décrite que par son producteur est symptomatique de l'approche Cartésienne de la notion de sens, qui, nous le verrons, n'est pas ce qui fonctionne le mieux en termes de données liées et web de données, bien au contraire.[151]

L'identification en LED

Ici, plus que jamais, l'ensemble des entités doivent être identifiées par des URI. L'unicité des URI reste un élément central du système. Ils peuvent même, dans certains cas, supplanter l'utilisation de référentiels communs[152]. Le schéma de l'identification doit rendre compte des liens et dépendances entre les niveaux de granularité, et doit être extensible. Le choix des URI http ici permettrait de garantir que les identifiants soient actionnables avec de simples outils web, et présenter une nouvelle source de possibilité pour le *data management* au sein de l'entreprise.

Le LED nous apprend donc ici une chose essentielle : l'homogénéisation n'est pas forcément un prérequis qu'il faut à tout prix réaliser. L'idée serait de pouvoir relier des données, métadonnées, ontologies très hétérogènes qui potentiellement se recoupent les unes les autres, tout en exploitant les spécificités de chacun et en créant des passerelles d'équivalence. Même au niveau des identifiants, cela peut s'avérer utile : cet idéal utopique de vouloir trouver un identifiant parfait, qui satisferait toutes les exigences, ne serait même peut-être pas une bonne chose, puisqu'il signifierait l'abandon de spécialisations qui, en entreprise en tout cas, conditionnent l'efficacité de celles-ci dans certains domaines métier.

---

[147] *Ibid.*

[148] DOCTOROW, Cory. *Metacrap: Putting the torch to seven straw-men of the meta-utopia*. 2001. [en ligne] Disponible sur https://www.well.com/~doctorow/metacrap.htm [consulté le 06/07/2017]

[149] LEGG, Catherine. *Op. Cit.*

[150] DOCTOROW, Cory. *Op. Cit.*

[151] Nous pourrons nous reporter à la partie 3 de ce mémoire pour développer ces arguments.

[152] BERMÈS, Emmanuelle, ISAAC, Antoine, POUPEAU, Gautier. *Op. Cit.*

Conclusion de la partie 1

Cette partie aura donc exploité plusieurs points qui aboutissent sur les constatations suivantes :

- Les identifiants, ou devrions nous dire, les méthodes d'identification, sont clairement la pierre angulaire de tous les systèmes de gestion des connaissances, quels qu'ils soient, car les entités ont besoin d'être nommées, localisées, contextualisée, pour ne serait-ce qu'exister,
- Ils prennent des formes différentes et sont souvent responsables de la réussite et de la praticité d'un système, puisqu'ils touchent à l'accès et à l'existence même des données sur le réseau,
- Les institutions culturelles et les entreprises ont des valeurs et des objectifs différents mais les moyens de les atteindre peuvent se recouper et s'enrichir l'un l'autre, (c'est pourquoi il est intéressant de développer des études de cas sur les deux types de structure),
- Les modèles de données et outils sont divers, complexes, mais de ce fait adaptables aux besoins précis des organisations. Ils évoluent rapidement et doivent en conséquence faire l'objet d'une veille sérieuse à ce sujet,
- Il est possible de créer des systèmes hybridant différentes méthodes, outils, solutions qui permettent de réussir cette adaptation et construire des systèmes personnalisés,
- Les identifiants et leur gestion permettent en outre de prendre la « température » des systèmes et d'évaluer leur réussite potentielle au niveau global, car les soucis surgissent toujours à l'accroissement du nombre de données gérées.

Avec ces notions, nous sommes mieux préparés à aborder les aspects techniques que nous verrons dans la partie 2, ainsi que les études de cas et recommandations et bonnes pratiques rassemblées dans la partie 3 de ce mémoire. Si le panorama est dense, il permettra au moins de comprendre l'étendue des possibilités qu'offrent les données liées et le web.

2. L'IDENTIFIANT SOUS TOUTES LES COUTURES

2.1. ANATOMIE DE L'IDENTIFIANT

2.1.1. Les 8 caractéristiques de l'identifiant idéal

Nous l'avons vu, la valeur d'un identifiant et son efficacité sont des problématiques qui divisent depuis longtemps les communautés professionnelles. Nous allons ici faire un récapitulatif des 8 caractéristiques de l'identifiant idéal, celui que nous imaginons pouvoir convenir à toutes les structures : étant aussi approprié aux besoins d'une petite bibliothèque de 3 employés qu'à ceux d'une firme multinationale important des produits alimentaires.

Cette liste est globalement un condensat des recherches effectuées dans le domaine, à partir des préconisations des professionnels contenues dans notre bibliographie. Elle n'est pas exhaustive, nous développerons ici chacune de ces caractéristiques (ou fonctionnalités) car elles semblent indispensables dans le cas de notre identifiant idéal. Evidemment, cela est utopique, mais permettra au moins de voir exactement quelles sont les caractéristiques globales d'un identifiant, son rôle précis, et quels critères il convient de prendre réellement en compte dans le choix d'un système, afin d'avoir une vue d'ensemble sur la praticité concrète du futur identifiant utilisé.

Unicité, pérennité, échelle

Tout d'abord, nous allons évoquer les 3 notions qui sont couramment les plus mises en valeur dans cet ensemble de qualités. La notion d'unicité est celle qui semble la plus évidente, bien qu'elle puisse présenter des nuances suivant les besoins. En informatique, l'unicité est primordiale. Une entité qui n'est pas identifiée de manière unique n'est pas déréférençable ni accessible, elle n'existe pas pour le système, on ne peut la distinguer des autres. Et dans ces environnements de communication pure, (contrairement à la réalité physique où l'on peut « être » sans être connu, la fameuse déclaration *je « pense donc je suis »* de Descartes), être localisé, c'est ce qui conditionne l'existence. C'est donc la question la plus incontournable en terme d'identification. Il s'agit donc ici de désigner une seule entité (un seul document ou un seul concept, personne, lieu, etc.) séparément des autres. Ici, la problématique se noue : doit-on identifier un document global qui engloberait l'ensemble des versions de ce même document? Ou bien justement l'identification devrait-elle se cantonner à discriminer les versions entre-elles uniquement en étant liées à un concept global? Nous revenons ici aux questions évoquées dans la partie précédente[153] à savoir qu'est-ce qui est identifié réellement. L'idée est de pouvoir référer à une entité, et une seule.

L'unicité de l'identifiant même, séparément de son référent, est tout aussi importante. Quel intérêt y aurait-il à avoir des identifiants qui se recoupent et se confondent ? La création automatique et aléatoire d'identifiants assure un certain degré d'unicité qui lui permet d'être univoque. Lorsque ce n'est pas le cas, il est, par conséquent, ambigu, car un même identifiant peut être relié à deux ressources différentes. On appelle cet écueil la « collision »*, et il est vital d'éviter ce genre de problème au sein d'un système d'identification. S'il est donc impossible (ni même souhaitable) d'avoir plusieurs référents pour un identifiant, l'inverse n'est pas vrai. Nous pourrions en effet trouver des entités qui sont les référentes de plusieurs

[153] Voir partie 1.2.1

identifiants. Nous verrons dans les cas pratiques de la partie 3 dans quelles conditions cela peut être bénéfique. En outre, on parle d'identifiant globalement unique lorsque cette unicité se vérifie sur l'ensemble des ressources, internes et externes à la structure. Cela suppose nécessairement une organisation centralisée au niveau international.

La problématique de la pérennité est une autre notion très controversée dans le domaine qui nous concerne. Qu'est-ce qu'être pérenne ? Cela implique nécessairement une conception du temps très subjective et cela alimente énormément les débats sur le sujet. Il s'agit d'une garantie de stabilité, qui permet un accès aux entités avec le même identifiant sur un temps « long » (là encore, on voit bien que cela peut être soumis à interprétations variables). Cette garantie est très importante car elle constitue une grande partie du retour sur investissement de la mise en place du système, et aussi sa qualité. Quel est l'objectif d'un système d'identification si ce n'est de faire en sorte qu'il soit utilisé sur la durée ? Ici nous pouvons considérer que les organisations appréhendent le temps à l'échelle de la vie de la structure, et peuvent ainsi trouver « suffisamment longue » une période de 30 ans par exemple. Là encore, cela dépend donc de l'objectif de chacune : est-ce une institution dite « de la mémoire » qui a pour but de conserver des documents à l'infini du mieux qu'ils peuvent l'être (BnF, archives…) ou est-ce une entreprise qui visualise son plan de développement sur 30 ans ?

Si initialement beaucoup considéraient le problème d'un point de vue technique, en cherchant des solutions au niveau logiciel et matériel pour pérenniser les accès et le déréférencement de manière générale, il s'est progressivement développé dans la communauté un consensus autour de la constatation que la pérennité n'est pas un problème technique, mais un problème de gouvernance[154]. Ainsi, l'indépendance vis-à-vis de la technique serait une des meilleures façons de garantir la pérennité, la technique et les solutions matérielles étant très changeantes et évolutives. De plus, la pérennité est dépendante de l'ensemble de la chaîne, tous les acteurs doivent donc être impliqués au même degré dans cette pérennité pour garantir une quelconque efficacité. C'est donc une caractéristique complexe à prendre en compte dans notre identification idéale, qui apporte beaucoup plus de questions que de réponses.

Enfin, l'échelle est ce qui permet d'ajuster l'une ou l'autre de ces précédentes notions. Souhaite-on avoir des identifiants uniques au sein d'un système ou uniques au niveau global ? Recherche-t-on la pérennité « éternelle » ou « suffisamment longue »[155] ? De même, souhaite-t-on garantir un accès local, global ou conceptuel ? La définition précise et explicite de ces paramètres d'échelle doit être liée aux besoins précis d'une structure, et détermine toute l'efficacité du système.

Granularité, Adaptabilité, accessibilité

La granularité est une caractéristique qui est très liée à la notion d'unicité. En effet, c'est ici que nous allons choisir ce que l'on identifie vraiment, et à quel niveau de détail nous allons nous limiter pour l'attribution des identifiants. Identifierons-nous chacun des différents formats de versions de chaque document web ? Allons-nous avoir des identifiants hiérarchiquement supérieurs identifiant des ensembles, tels une collection ou un cadre de structure ? L'implémentation d'un niveau très haut

[154] Ministère de la Culture et de la Communication. *Op. Cit.*

[155] *Ibid.*

de granularité peut être chronophage et les possibilités vont de l'infiniment grand à l'infiniment petit. Il s'agit donc de définir des règles qui correspondent aux besoins concrets d'utilisation en fonction, là encore, de la structure, et de voir les probabilités d'évolution des besoins également sur le long terme.

L'adaptabilité correspond à la capacité d'un système à englober des identifiants extérieurs et de versions précédentes ou de justement se modifier en fonction de l'évolution de ces besoins ou de l'environnement. Il peut préexister dans la structure des modèles d'identification dont il faut assurer la rétroactivité, comme il est intéressant de pouvoir intégrer des identifiants physiques déjà utilisés dans le domaine. Cela présente plusieurs avantages : tout d'abord cela permet d'assurer la continuité d'un service et de ne pas générer de frustration pour les usagers (frustration qui, nous l'avons vu, est très nocive pour la popularité d'un service et son accueil par les utilisateurs) ; ensuite cela permet d'accroitre les possibilités d'interopérabilité* (la réutilisation d'identifiants physiques ou globaux dans un secteur permet de favoriser l'appréhension globale du système, les liens avec les autres ainsi que l'utilisation d'un « langage » commun) ; enfin cela participe à réduire les temps d'implémentation, car l'élaboration d'une base conceptuelle personnalisée complète nécessite de gros efforts d'analyse, de tests, de définition de standards, etc. En effet, il suffirait d'intégrer les identifiants déjà à disposition pour bénéficier de systèmes qui fonctionnent pour le type d'entité en question.

> « La pérennité repose en outre sur la capacité à s'adapter aux changements de l'environnement, et il est nécessaire de pouvoir étendre les identifiants et les adapter au fur et à mesure de l'apparition de nouvelles ressources, des évolutions du réseau, des standards du web, des capacités des navigateurs. »[156]

Enfin, l'accessibilité est une notion qui pose d'autres questions importantes, comme nous l'avons vu avec les problématiques liées à l'URN et l'URL, l'articulation des nécessités pour l'identifiant de mêler dénomination, contexte et topographie. L'accessibilité est liée aux méthodes d'accès à la ressource, autrement dit la capacité d'un identifiant à être actionnable. Pour Tim Berners-Lee, l'identifiant idéal est actionnable. Il préconise de ce fait fortement l'utilisation d'URI http, critère qui fait d'ailleurs partie des caractéristiques primordiales de l'identifiant. Pour Danny Ayers et Max Völkel, du W3C, il faut qu'il y ait une « *description de la ressource identifiée qui puisse être retrouvée avec les technologies standard du web.* »[157] Ainsi, dans le cas où il est nécessaire de mettre en place une actionnabilité des identifiants non basés sur le protocole http, nous devons systématiquement mettre en place ce qu'on appelle des résolveurs*, qui sont des *plug-ins* de navigateurs ou services qui permettent ce déréférencement. Ce n'est cependant qu'un détournement habile qui finalement utilise tout de même http. Il peut être interne à une structure ou externe, géré indépendamment par une autorité. Cela nécessite tout de même une dépendance technique, qui ne fait que contourner la facilité et la praticité du déréférencement classique via le web.

[156] *Ibid.*

[157] AYERS, Danny, VÖLKEL, Max. *Op. Cit.*

Citabilité, universalité

Ici nous allons évoquer deux caractéristiques supplémentaires, qui recoupent en partie celles précédemment évoquées, mais qui doivent être citées de par leur présence dans le vocabulaire professionnel et leurs spécificités propres.

La caractéristique de *citabilité* a été introduite par Emmanuelle Bermès dans son document sur les identifiants pérennes[158]. Celle-ci définit le besoin de pouvoir citer une ressource, un *web document*, une entité via les canaux numériques à travers une référence stable, permanente, permettant de nommer et de retrouver la ressource. C'est donc un concept à cheval sur plusieurs des précédents évoqués qui sont l'unicité, la pérennité, et l'accessibilité. Mais il apporte également une autre caractéristique importante, celle du contexte. Pour qu'une référence soit réellement *citable*, celle-ci doit en effet induire un contexte : l'humain qui cite doit pouvoir comprendre ce qu'il cite et en tirer du sens. Si ce contexte n'est pas indiqué clairement dans la dénomination concrète de l'identifiant (identifiant opaques, par exemple, nous développerons cette idée plus avant), d'autres moyens doivent être employés pour permettre la manipulation des informations. Ainsi, des métadonnées doivent systématiquement être ajoutées lors de l'enregistrement de l'identifiant d'une ressource (tels que dans les systèmes d'identification DOI ou encore ARK) et doivent se trouver dans un système permettant l'échange et le transfert de celles-ci. Elles doivent en outre elles-mêmes être adaptables et présenter une cohérence, même dans le cas d'une grande collection hétérogène.

Venons-en à présent à la dernière caractéristique, l'universalité. Celle-ci pourrait d'ailleurs tout aussi bien être un objectif (conséquence de la conjonction de plusieurs caractéristiques) qu'un critère à part entière. L'universalité, notion chère à Paul Otlet dans son *Traité de la documentation* est en effet un concept que l'on pourrait tout à fait évoquer dans la construction de notre identifiant idéal. Elle recoupe cette idée d'identifiant unique global, c'est-à-dire permettant de tout identifier et d'être utilisé dans toutes les circonstances, et l'idée de l'accessibilité, c'est-à-dire un identifiant accessible en permanence et par tous. Tout le monde doit pouvoir identifier le document, sans contrainte d'appartenance à une organisation.[159] Cette notion est intéressante car elle cristallise le dilemme entre communication et conservation : si la conservation au sens archivistique du terme nécessite un accès aux ressources restreint pour en limiter l'utilisation, il semblerait que, sur le web, la pérennisation se fait justement par l'utilisation. Ce qui est utile est par conséquent maintenu, ce qui ne l'est pas est laissé à l'abandon. C'est une sorte de sélection naturelle appliquée au web. En théorie, donc, l'universalité de l'identifiant, permettant à n'importe qui depuis un poste de travail de comprendre et d'accéder à une entité, est plutôt bénéfique pour l'ensemble et serait un bon garant de sa pérennité. Cependant s'additionne le problème des droits d'accès, qui est très complexe à gérer au vu des nombreux paramètres à prendre en compte, et de plus cela n'enlève rien à la question des données d'entreprise internes avec leurs conditions de confidentialité. Toutefois, le concept reste intéressant, car il apporte le lien qui manquait avec les données liées que nous avions développées dans la première partie de ce mémoire : l'objectif d'universalité d'un identifiant permet de l'ancrer dans un graphe global de données qui en développe intrinsèquement toutes les caractéristiques précédemment citées.

---

[158] Ministère de la Culture et de la Communication. *Op. Cit.*

[159] ARCHIMBAUD, Jean-Luc. *Op. Cit.*

2.1.2. Les systèmes d'identifiants pérennes

De quoi s'agit-t-il

Qu'est-ce qu'un « identifiant pérenne » ?

> « Un identifiant pérenne est une chaîne de caractères alphanumériques qui a pour fonction d'identifier de manière stable un document, une ressource ou une entité quelle que soit sa nature. »[160]

Les systèmes d'identifiants pérennes, ou PID (*Persistent ID*) sont donc un peu nos identifiants idéaux décrits précédemment : ils sont uniques, adaptables, et interprétables soit directement, soit au moyen d'un résolveur. Ce sont les bons élèves de notre petit groupe d'identifiants, ils se veulent globaux et bien construits afin de garantir la pérennité pour peu que la structure elle-même remplisse ces prérequis. Ils correspondent également à cet idéal d'universalité, même si pour réellement s'y conformer complètement ils devraient être totalement gratuits (au niveau de l'attribution de l'identifiant et de sa maintenance notamment, tout le monde devrait pouvoir en créer) et libre d'accès (accessible par tous sans conditions de paiement ou de statut). L'entre-deux choisi pour la plupart des situations concernant la problématique de droits est de restreindre l'accès concret à la ressource sans pour autant interdire l'appréhension du référent. Nous pouvons donc, en théorie, savoir à quoi correspond un identifiant de PID même si nous ne pouvons pas y avoir concrètement accès.

Les PID ont une syntaxe commune composée d'une structure générale et des propriétés à valeur spécifique. La syntaxe est généralement basée sur les spécifications du W3C [161]: ils possèdent un préfixe qui indique le contexte dans lequel l'identifiant est attribué, suivi d'un élément de désignation de l'autorité nommante, puis du nom de la ressource. La globalité s'obtient par l'ajout d'un élément globalement unique, tel un code d'institution nommante. Ces codes doivent être gérés par un organisme ou bien par une association d'organismes qui souhaitent s'accorder sur un annuaire commun. Ils ont donc en charge la gestion de la syntaxe des URI et des modèles organisationnels plus ou moins centralisés et parfois ils fournissent des logiciels « clés en main ».[162] Parmi ceux-ci figurent les renommés DOI et ARK, repérés parmi les pionniers du genre.

Des questions de citabilité, sémantique et opacité

La problématique de l'opacité et de la signifiance dans l'identification est intéressante. Elle pointe assez bien la différence que présente la gestion humaine, contextuelle, sémantique et celle, « stupide »[163] mais très efficace des ordinateurs. Nous avons d'un côté un identifiant qui est signifiant : c'est-à-dire que dans sa constitution se trouvent des mots intelligibles par un humain, du type : un titre, un nom d'auteur, une date d'enregistrement pour un ouvrage par exemple. De l'autre, nous avons un identifiant opaque, c'est-à-dire une suite de caractères

[160] Ministère de la Culture et de la Communication. *Op. Cit.*

[161] *Ibid.*

[162] *Ibid.*

[163] En référence aux *Information stupids*, les machines telles qu'évoquées par Mirna Willer et Gordon Dunsire dans *Bibliographic Information Organization in the Semantic Web*, Chandos Publishing, 2013. par opposition aux humains.

alphanumériques auquel un individu ne trouverait absolument rien de comparable dans sa mémoire et ne pourrait pas créer de lien apportant du sens.

Le problème de la signifiance, c'est qu'elle embarque/intègre (*embedded*) dans son identification du sens, et par définition le sens peut « pourrir »[164]. Il évolue, change, il est fluide comme le langage et un mot tout à fait anodin à une certaine époque veut dire totalement autre chose un siècle plus tard.[165] Il est également soumis à l'ambiguïté et la désuétude des termes, car il est intimement lié à son référent. En fonction des besoins, un identifiant signifiant est très commode : il est manipulable et appréhendable par des humains. Mais dans le cas d'un fonds d'entités extrêmement hétérogène, cela pose un gros problème puisqu'il n'y aura pas d'unicité (au sens esthétique du terme) dans le système.

Les identifiants opaques présentent quant à eux beaucoup d'avantages mais sont plus difficilement gérables par un humain. Ils nécessitent de surcroit la maintenance d'un système de référence distinct implémenté techniquement, donc potentiellement plus couteux et soumis aux évolutions technologiques. Mais ils sont dissociés de l'évolution d'une ressource et de sa description, et sont plus facilement uniques et pérennes. La notion d'affordance* est liée à cette problématique, elle correspond à une situation dans laquelle les caractéristiques d'un objet correspondent à ses fonctionnalités et son usage. Dans notre cas, cela réfère à la capacité que l'on a de générer un identifiant syntaxiquement correct à partir d'un contenu.[166] Les identifiants opaques ont donc, logiquement, un niveau d'affordance très faible. De même, si en pratique les identifiants signifiants devraient avoir un niveau d'affordance élevé, ce n'est pas toujours le cas, cela dépend surtout de la fiabilité de l'attributaire (est-ce qu'il identifie bien ce qu'il prétend identifier ?). On ne peut pas concrètement remonter à la source du processus ni même au référent en observant la chaîne de caractère qui l'identifie. Cela est à la fois un bien et un mal. Mais alors, que choisir entre opaque ou signifiant ?

Bide et Green, dans leur document de 1999 sur les identifiants uniques considèrent que « *trouver une solution à cette problématique a autant de chances d'aboutir que la recherche de la licorne.* »[167] Le ton est donné. Ils imaginent que l'on devra dans un premier temps commencer par l'attribution d'identifiants signifiants qui finiront par être remplacés à terme par des identifiants opaques. En pratique et avec le recul actuel, cette problématique a été gérée à l'aide d'une approche hybride, dans les PID, par l'utilisation d'identifiants opaques aux « modules facultatifs » extensibles et par la gestion de métadonnées. Nous développerons cet aspect lors de l'exploration des identifiants ARK.

L'identifiant, un être intelligent qui communique avec ses pairs

Ainsi, il est possible d'incorporer de la signifiance dans un identifiant pérenne, mais cela doit être fait de manière réfléchie. Le terme anglo-saxon pour la signifiance est littéralement *l'intelligence.* Ce concept est très lié au type

[164] Cela fait référence au phénomène du *link rot* couramment utilisé dans le langage professionnel, notamment par John A. Kunze.

[165] Nous citerons notamment à ce propos l'exemple du mot « Gay » en anglais, tiré de l'article de John A. Kunze *Towards Electronic Persistence Using ARK Identifiers*, publié en 2003 qui a un sens bien différent actuellement (se référant à l'homosexualité) de celui utilisé au siècle dernier (se référant à un état joyeux, enjoué).

[166] PASKIN, Norman, RUST, Godfrey. *Op. Cit.*

[167] GREEN, Brian, BIDE, Mark. *Unique Identifiers: a brief introduction*. 1999. 11p. [en ligne] Disponible sur http://www.bic.org.uk/files/pdfs/uniquid.pdf [consulté le 29/11/2016]

d'intelligence contextuelle typiquement humaine que l'on essaye d'inculquer aux machines par l'intelligence artificielle, principe même du web sémantique et au cœur de notre étude. Nous distinguons donc trois types d'intelligence (de signifiance) au sein de l'identifiant :

- l'intelligence vitale,
- l'intelligence risquée,
- l'intelligence dynamique.[168]

L'intelligence vitale correspond au processus d'attribution, on la retrouve dans toutes les URI : elle est souvent contenue dans les préfixes permettant le déréférencement ou la liaison à l'attributaire (http ou ARK, par exemple), elle peut être transparente ou cryptée (par exemple : le code à 3 chiffres 978 au début d'un EAN qui signifie que la suite correspond à un ISBN).[169] L'intelligence risquée est quant à elle l'information qui renseigne sur l'attributaire ou la date d'attribution. Elle est parfois sujette à quiproquo : l'institution identifiée est-elle l'attributaire de l'identifiant ou celle qui donne accès à la ressource ? Norman Paskin et Geoffrey Rust considèrent que toute information pérenne devrait être, non pas incluse dans l'identifiant même, mais plutôt dans les métadonnées. Ils nomment en outre les identifiants n'embarquant pas d'attributs les identifiants de première classe, c'est dire s'ils insistent pour que l'intelligence soit utilisée avec parcimonie dans l'identification. Enfin, l'intelligence dynamique telle qu'ils la définissent est la pire intelligence possible à intégrer dans un identifiant : elle concerne l'ajout de données à caractère strictement technique faisant lien à l'implémentation. Au-delà d'être extrêmement changeantes, évolutives et conditionnant une obsolescence programmée certaine, elles peuvent délivrer des informations que l'on ne souhaiterait pas voir diffusées. La solution viendrait donc de l'emploi de métadonnées pour apporter les informations signifiantes non comprises initialement dans le corps de l'identifiant.

L'efficacité et l'utilité d'un système d'identifiant pérenne tient également dans son rapport avec les autres, car non, même si ses concepteurs le voudraient très chèrement, il n'est pas souvent seul sur le marché de l'identification globale. L'interopérabilité des systèmes d'identifiants pérennes est donc cruciale, ne serait-ce que pour convenir à cette identification idéale qui se doit de tout prendre en compte de manière globale. L'on peut avoir 3 types d'interopérabilité. Tout d'abord, une interopérabilité syntaxique : le système reconnait une certaine syntaxe dans une chaîne de caractère (identifiant) et lance des actions correspondantes ; ensuite, une interopérabilité sémantique : le système sait si l'on se réfère ou pas à la même entité grâce à une cartographique des entités alternatives (par exemple un alignement des modèles de données comme nous avons pu l'évoquer précédemment) ; et enfin une interopérabilité de communauté : le système intègre la gestion de droits pour la restriction d'usage des données, partant du principe qu'une organisation souhaitant mettre en ligne ses données est *de facto* d'accord pour qu'elles soient en *open-access*.[170]

Le choix d'utiliser des identifiants provenant d'un système d'identification pérenne est donc une manifestation assez intéressante de cette volonté d'identification globale idéale que nous avions évoquée précédemment. Cette

[168] PASKIN, Norman, RUST, Godfrey. *Op. Cit.*

[169] *Ibid.*

[170] *Ibid.*

constatation permet de voir où en sont les avancées théoriques et technologiques sur le sujet et représente des cas d'usages parlants. Nous allons à présent nous focaliser sur le web principalement et sur la manière de gérer l'identification lorsqu'elle n'est pas régulée par des PID.

2.1.3. Méthodes d'identification pour les données sur le web

Les données, nous l'avons vu, ont besoin d'être identifiées sur le web pour exister et être utilisées. Cela se fait systématiquement via la construction d'URL spécifiques à chaque document web. Mais ceux-ci ne devraient pas obligatoirement être aléatoires, ils ont un sens et peuvent être améliorés. Ici nous évoquons les *web document* dans leur ensemble, c'est-à-dire les entités qui sont déréférençables sur le web, qui ont une « palpabilité » numérique (ou justement devrait-on dire à l'inverse, une dématérialisation numérique) et qui sont soit des entités à part entière, soit des représentations d'entités réelles.

Le web, un environnement plus restrictif qu'on ne le croit

Les pratiques d'attribution d'URL sur le web sont un capharnaüm sans nom. Chacun y va de son expertise, créant des URL à foison dans toutes les formes possibles et imaginables.[171] De plus, jusqu'à 2010-2011 se sont développées des méthodes de *Search Engine Optimization* (SEO) qui profitent des failles présentes dans les algorithmes de *PageRank** des moteurs de recherche afin de générer du trafic fictif (au moyen de doublons, de liens, etc.) et ainsi d'être mieux cotés.

Cependant, en 2012 Google développe et met en service un algorithme de moissonnage particulier qui va bouleverser la vision que l'on avait de la gestion web des ressources. En effet, il va traquer les doublons et les sites proposant trop de liens morts (erreurs 404) afin de les faire descendre dans les listes des résultats et ainsi fournir une expérience utilisateur plus agréable. Cette chasse aux techniques de référencement abusives, sous le nom de Panda (puis par la suite sa mise à jour Penguin), encourage désormais les fournisseurs de contenus sérieux à ne pas maintenir d'identifiants pérennes dont le contenu serait supprimé. Cela oblige ces derniers, pour garder leur positionnement dans les réponses du moteur (et ainsi conditionner leur utilisation), à effectuer systématiquement des redirections pour le maintien de leurs URL (code statut* http 303) lorsque le contenu change et traquer les doublons.

L'efficacité du web comme outil global de gestion de données gagne donc à voir s'homogénéiser ou du moins voir interopérer les pratiques d'attribution d'identifiants de manière générale : pour les fournisseurs d'une part car ils sont récompensés de la qualité du service, et pour les utilisateurs d'autre part qui ont ainsi une expérience plus efficace lors de leurs navigations. Certains professionnels aux W3C ont fait de ce combat au quotidien un corps de métier.

Construction de l'identifiant de document web

Les meilleurs identifiants pour les documents web ont un design simple, sont stables, facilement actionnables.[172] La définition d'un identifiant pour ce type de ressource nécessite le choix d'un nom de domaine et d'un formalisme. Comme

[171] C'est toutefois l'opinion partagée par nombre de professionnels du domaine, notamment Ted Nelson ou Tim Berners-Lee.

[172] AYERS, Danny, VÖLKEL, Max. *Op. Cit.*

évoqué précédemment, les éléments introduits dans cette construction ne doivent pas comprendre des informations touchant au contexte technique ou à la localisation de la ressource. L'utilisation de PID peut être un plus mais cela doit présenter un intérêt en fonction des besoins précis des structures (par exemple pour de gros fournisseurs de contenus qui ont un objectif de pérennité).

De manière générale, il ressort que laisser un logiciel s'occuper seul de l'attribution des identifiants est une très mauvaise idée.[173] En effet, un logiciel sera plus enclin à introduire un certain nombre de détails techniques superflus et gênants, et les futures migrations risquent d'être « douloureuses ». L'identification doit se réfléchir en amont, se construire selon un besoin spécifique indépendamment de l'implémentation technique. En effet, l'indépendance des URL rend possible le déplacement des ressources sans modifier les URL (c'est très intéressant dans le cas que nous évoquions plus haut avec le besoin d'indépendance des URL pour faciliter les redirections), et l'on peut également complètement bouleverser l'implémentation sans devoir les changer, ce qui apporte le facteur de stabilité recherché. Ici, on gère des documents présents numériquement, mais la question ne se pose pas concrètement avec les objets réels.

Seth Van Hooland et Max Verborgh évoquent deux types d'excuses pour ne pas utiliser des URL : soit l'institution n'a plus la main sur le nom de domaine, auquel cas cela rend difficile la définition d'URL personnalisées ; soit la ressource identifiée n'existe plus, auquel cas la réutilisation de l'URL amènerait des confusions néfastes.[174] Dans le cas d'attribution d'URL signifiantes, tout l'enjeu réside dans la recherche de propriétés invariables qui concernent la ressource. Par exemple, le VIAF (*Virtual Authority File*) associe à des artistes un numéro unique qui sert à construire l'URL pour chacune des œuvres. Côté Wikipédia, l'URL récupère la dénomination complète et ajoute des éléments de désambiguïsation.[175] En outre, l'utilisation d'indices sur la hiérarchie des ressources (par exemple : personne/acteur/Samuel_L_Jackson) créé une attente au niveau de l'usager de liste plus générales, mais cela peut prêter à confusion vis-à-vis de la corrélation qui peut se faire avec la hiérarchie des systèmes de fichiers qui, réellement n'ont rien à voir avec les processus d'identification.[176]

La liaison entre les différentes représentations d'un document est aussi importante : Danny Ayers et Max Völkel[177] recommandent explicitement que ces versions soient liées entre elles. Par exemple, dans les *header** des documents html, devraient se trouver en lien l'accès à la version RDF de ce même document. Dans le cas où cette représentation en RDF n'est pas la représentation exacte du document, le lien devrait être fait grâce à des déclarations de relation/équivalence. C'est une façon de compléter un graphe de données à l'échelle du web. La question se pose tout de même, est-ce la pérennité du format de la ressource qui est nécessaire ou uniquement son contenu intellectuel ?[178]

---

[173] VAN HOOLAND, Seth. VERBORGH, Ruben. *Op. Cit.*

[174] *Ibid.*

[175] *Ibid.*

[176] Cette notion est développée plus longuement dans l'introduction de ce mémoire.

[177] AYERS, Danny, VÖLKEL, Max. *Op. Cit.*

[178] Cette question a été posée en 2004 lors du Séminaire ERPA à Cork au sujet des identifiants pérennes. ERPA Seminar. *Persistent Identifiers, Final Report*. Cork, Ireland. 17-18 June 2004. Séminaire Erpanet, [en ligne] Disponible sur http://www.erpanet.org/events/2004/cork/Cork%20Report.pdf [consulté le 12/04/2017]

Au sujet de l'attribution de métadonnées

Cependant, il est important de rappeler que lorsque l'on produit une représentation d'une entité, certains aspects de son contexte sont inévitablement perdus. Nous l'avons vu, lors de l'attribution d'identifiants signifiants et, à plus forte raison, opaques, l'ajout de métadonnées permet de compléter la contextualisation des données que l'on publie et d'en augmenter la portée, tant au niveau du référencement que dans les méthodes utilisées pour les exploiter. La métadonnée de qualité est donc un élément clé de l'enrichissement de la ressource qui permet à l'utilisation d'être optimisée. Mais que trouve-t-on sur le web comme métadonnées disponibles pour notre usage ?

Mirna Willer et Gordon Dunsire définissent trois types de producteurs de métadonnées sur le web [179]:

- Les professionnels de l'information, qui produisent des métadonnées efficaces sur un panel très large de type de ressources, pour une majorité d'usagers, mais en quantité très faible,
- Les « amateurs », constitués pour la plupart du grand public surfant sur Internet et publiant du contenu, produisant des métadonnées plus spécifiques et personnelles sur des sujets restreints, de quantité moyenne,
- Les « *Information stupids* » ou machines, qui produisent des métadonnées automatiquement, de manière « bête et méchante », en très gros volume.

Avec cette catégorisation, l'important est de savoir à quoi l'on a affaire : sont-ce des métadonnées produites par un professionnel, un amateur ou un *information stupid* ? La provenance est importante, les professionnels de l'information produisant des métadonnées en petite quantité mais de très bonne qualité, tandis que les deux autres fournissent plus de données mais de qualité douteuse. Evidemment, la question à se poser est la suivante : vaut-il mieux avoir peu (voire presque pas) de métadonnées de très bonne qualité ou de la métadonnée en plus grand nombre de moindre qualité ? Nous développerons ces questionnements dans la troisième partie de ce mémoire.

2.1.4. Méthodes d'identification pour les objets réels

Le cœur du problème : représentation versus description

Comme nous l'avons vu en partie 1, comprendre ce qu'une URI désigne réellement est probablement la question la plus complexe à laquelle répondre actuellement. Le débat dans la communauté professionnelle est de savoir si l'URI doit être comprise comme une référence à un document ou bien comme une référence à une ressource.

Nous sommes d'accord, chaque URI pointant sur un document web soumise à un http GET doit répondre par le code statut 200. Mais un problème survient lors du besoin pour le web sémantique d'identifier des objets réels. Le protocole http ne peut pas « aller chercher » cet objet réel. Par exemple, on peut associer à une personne un numéro de téléphone, une adresse e-mail, une photo, mais on ne peut pas concrètement « l'invoquer » sur le web. Ainsi, la commande GET du *header* généralement utilisée dans ce cas-là ne peut pas être satisfaite et cela créé un blocage

[179] WILLER, Mirna, DUNSIRE, Gordon. *Op. Cit.*

que l'on connait comme le *http Range 14*[180], que nous avions évoqué en première partie. Mais alors, comment utiliser les URI pour des entités qui n'ont pas d'existence numérique ?

Ici se cristallise la délimitation entre représentation et description. Pour savoir à quoi chaque URI fait référence, il faudrait pouvoir regarder dans un moteur de recherche une petite description de l'objet, à défaut d'en trouver une représentation. La question est de savoir où poser la limite entre deux cas : le cas où l'on a la possibilité de voir l'objet lui-même, et le cas où seuls des éléments de description sont disponibles. Ainsi, si on fait une requête http GET sur l'URI d'un toucan on doit pouvoir récupérer un document qui décrit le toucan, tout en faisant en sorte qu'il soit explicite que ce n'est pas la représentation du toucan lui-même ; et si on fait une requête http GET sur l'URI du document à propos du toucan on doit pouvoir avoir une réponse avec le code statut 200, et celle-ci doit contenir des triplets RDF qui pointent et réfèrent à l'URI d'un toucan.[181]

Danny Ayers et Max Völkel du W3C présentent dans leur document *Cool URI's for the semantic web* datant de 2008 deux méthodes qui pourraient être utilisées dans l'appréhension de ces fameuses URI d'objets réels.

Le système des Hash URIs (URI dièse)

L'idée principale évoquée ici est d'utiliser des URI spécifiques pour les objets réels qui ne sont pas directement déréférençables. Celles-ci se distinguent par l'ajout d'un signe dièse (hashtag #) avant la partie qui détermine le concept, par exemple : www.site-merveilleux.com/apropos#moi. Ainsi, la personne souhaitant déréférencer cette URI amènera son navigateur à ne pas tenir compte de ce qui se tient derrière le signe dièse pour amener l'utilisateur à la page www.site-merveilleux.com/apropos. Sur cette page, l'idée serait de proposer via un double niveau de négociation de contenu des descriptions différentes en fonction du dièse sélectionné, tout en permettant également d'apporter des représentations différentes de cette description en RDF ou en HTML suivant les préférences du navigateur.[182]

Cependant, ce système présente quelques défauts, notamment le fait que le client doive télécharger l'entièreté des données des autres ressources (ce qui devient compliqué sur un très grand nombre de données, nous pouvons imaginer pourquoi). En sus, certaines API ne font pas la distinction avec les dièses, qui sont considérés comme des caractères semblables aux autres pouvant faire partie intégrante de l'URI. Cela pose également problème avec les systèmes de NER (*Named-Entity Recognition* que nous avions évoqués plus haut), s'ils prennent les URI d'objet réels pour des URI de documents cela peut créer beaucoup de confusion et des réattributions non souhaitées d'identifiants.[183] Cependant ce système reste intéressant pour les implémentations rapides en RDF et il évite les nombreuses requêtes http (typiques de la seconde méthode que nous allons aborder juste après).[184] De plus, une famille d'URI peut partager une même chaîne de caractère

[180] Les discussions autour de la question se trouvent répertoriées dans un fil d'e-mails au sein du W3C publiés ouvertement via https://www.w3.org/2001/tag/group/track/issues/14.

[181] DAVIS, Ian. *Is 303 really necessary?* Blog de Ian Davis, 2010. [en ligne] Disponible sur http://blog.iandavis.com/2010/11/04/is-303-really-necessary/ [consulté le 01/06/2017]

[182] AYERS, Danny, VÖLKEL, Max. *Op. Cit.*

[183] VAN HOOLAND, Seth. VERBORGH, Ruben. *Op. Cit.*

[184] *Ibid.*

avant le dièse, ce qui peut être avantageux dans le cas de recherches thématiques. Il est en outre facile de naviguer et de comprendre ce sur quoi l'on tombe lorsque l'on observe l'URI déréférencée.

La redirection 303, ou URI 303

Le principe de cette méthode est de réaliser une redirection systématique vers une description lors du déréférencement d'un identifiant d'objet réel. Etant donné qu'il n'y a pas de représentation appropriée à renvoyer, le client reçoit un statut code 303 qui va proposer une nouvelle URL (l'URL devant alors être l'objet d'une nouvelle requête http GET pour atteindre le *web document* en question), la négociation de contenu se faisant si besoin lors de cette nouvelle requête. Dans le cas où il y a plusieurs descriptions possibles, l'implémentation de la négociation de contenu peut se faire directement depuis l'identifiant de l'objet réel.

Pour mettre ce système en place, il faut donc configurer le serveur du fournisseur pour qu'il renvoie systématiquement des réponses 303 aux requêtes concernant les URL d'objets réels. Cette solution de redirection est plus flexible que la précédente, car chaque cible d'URL peut être configurée différemment pour chaque entité, permettant l'évolutivité du système. Dans son document *Is 303 really necessary ?* de 2010, Ian Davis considère que la solution de la redirection n'est pas la panacée. En effet, il avance des arguments intéressants [185]:

- elle demande énormément d'aller-retour entre usagers et serveur (et plusieurs requêtes) ce qui alourdit considérablement les échanges ;
- une seule description peut être vraiment liée à l'URI du toucan ;
- l'usager demande une URI spécifique mais se retrouve face à une autre, cela est créateur de confusion et d'ambiguïté[186] ;
- c'est plus compliqué à mettre en place ;
- cela ne peut pas être implémenté sur un serveur web statique ;
- lorsque le serveur web cesse d'exister certaines informations sont perdues ;
- la redirection 303 ne peut pas être utilisée pour d'autres utilisations, parce que là encore, cela créerait de la confusion ;
- et enfin l'opérateur du serveur doit pouvoir décider et comprendre lui-même quels sont les documents web et quels sont les objets réels.

Pour conclure, il suggère que le simple fait que l'on doive proposer ce système aux développeurs web ne fait que démontrer qu'il n'est pas la voie la plus évidente, et cela va jusqu'à desservir le web de données global en créant une résistance générale des développeurs envers l'implantation de ces mécanismes. L'auteur considère d'ailleurs que l'on devrait purement et simplement abandonner le système de redirection 303 pour ce type d'entités. Il propose en revanche une solution plus pragmatique, se délestant un peu de l'aspect philosophique lié au fond du problème *http range 14* : l'URI du toucan et l'URI du document seraient étroitement liées et leur déréférencement renverrait systématiquement à la même description, donc générerait le code statut 200. Celle-ci comprendrait dans ses métadonnées une propriété qui spécifierait le lien entre les deux URI par un « est décrit par » permettant de faire la distinction et d'identifier leur rapport mutuel. En pratique,

[185] DAVIS, Ian. *Op. Cit.*

[186] Et nous savons combien sur le web, nous n'aimons pas ça !

cette solution a été remise en cause de nombreuses fois et de nombreuses personnes se sont jointes à la discussion pour en débattre.

Les Archives de France, qui publient en ligne en *open-access* des ressources provenant des collections nationales et départementales, utilisent cependant la redirection 303 pour leur site Frances archives. C'est un système qui semble convenir à leurs besoins. En outre, il est possible de faire des combinaisons de différentes solutions pour s'adapter à une structure, par exemple une hybridation de la solution du Hash et de celle des URI 303. Cela peut même permettre d'avoir une plus grande finesse et flexibilité dans la gestion de ces URI d'objets réels, en séparant par petits fragments un grand volume de données[187].

Conclusion de la partie 2.1

Cette partie 2.1 nous a permis de rentrer dans les notions de base sur l'identification et de nous armer pour l'appréhension des systèmes, et nous allons à présent pouvoir aller spécifiquement sur les solutions proposées actuellement sur le marché.

[187] AYERS, Danny, VÖLKEL, Max. *Op. Cit.*

2.2. ETAT DES LIEUX DES SYSTEMES D'IDENTIFICATION

Dans cette sous-partie, nous allons explorer ce qui se fait en terme d'identifiants actuellement dans l'optique du partage sur le web de données. Nous laisserons de côté volontairement des identifiants qui ne correspondent pas à notre périmètre : notamment les systèmes d'identification attribués par automatisation (non lisibles par des humains) tels que RFID, les codes-barres… ; les systèmes se basant sur les empreintes numériques tels ContentID, PicScout, Soundmouse, Digimarc Attributor… ; et les systèmes d'identification technique trop spécifiques à un domaine (par exemple les identifications en sciences type médecine et physique, les identifiants de produits de consommation classiques, ou encore les codifications ISO correspondant à l'identification des pays, des territoires, des organismes bancaires ou des individus tels que le numéro de sécurité sociale). Nous n'irons pas non plus en détail sur les standards spécifiques à la normalisation des métadonnées en particulier (exemple : Avails, MovieLabs, DDEX) même si ceux-ci concernent des contenus intellectuels et parfois utilisent leurs propres identifiants.

Nous nous attacherons plutôt à des identifiants potentiellement utilisables et réutilisables, et verrons plus en profondeur un large choix d'identifiants de produits éditoriaux, des identifiants de contenus et de concepts, des identifiants d'agents, et des méthodes de gestion d'identifiant (comportant des identifiants, des protocoles et des exemples d'interopérabilité).

2.2.1. Les identifiants internationaux gérés par l'ISO

Ici nous allons nous concentrer sur quelques-unes des différentes normes issues de l'International Standard Organisation (ISO), et tout particulièrement à celles provenant du groupe ISO TC46 (et sa subdivision SC9 relative à l'identification et à la description), ayant trait à l'information et à la documentation..

ISBN

L'ISBN est l'acronyme de *International Standard Book Number*, et est un identifiant ISO (issu de l'*International Standard Organisation*) qui sert pour toutes les publications uniques. Il identifie la manifestation d'une œuvre, et non l'œuvre en tant qu'entité conceptuelle. Il fait partie intégrante des EAN (*European Article Numbering*) dans lesquels la publication en tant qu'article (produit) a pour EAN[188] la reprise intégrale de l'ISBN. L'attribution d'un préfixe d'éditeur est payante.

L'ISBN est composé de 13 chiffres (initialement 10, mais cela a évolué), sa syntaxe est la suivante[189] :

- Le premier segment, nommé G, correspond toujours à 978 lorsque ce qui suit concerne un ISBN10 (jusqu'à 2006), et 979 lorsque ce qui suit concerne un ISBN13 (à partir de 2007),
- Le deuxième segment, nommé A correspond à la zone géographique et peut être composé de 1 à 3 chiffres. Il existe des regroupements logiques (bien qu'il y ait techniquement la possibilité d'attribuer un numéro par pays) : notamment pour les pays francophones qui partagent le numéro 10 en ISBN10 et le numéro 2 en ISBN13,

[188] Voir page Wikipédia sur l'EAN13 [en ligne] disponible sur : https://fr.wikipedia.org/wiki/EAN_13 [consulté le 02/05/2017]

[189] ARCHIMBAUD, Jean-Luc. *Op. Cit.*

- Le troisième segment, nommé B correspond à un numéro d'éditeur qui a été attribué à une structure quelle qu'elle soit par l'une des agences d'enregistrement des ISBN (en France il s'agit de l'AFNIL). 1 à 7 chiffres le constituent,
- Le quatrième segment, nommé C correspond à un numéro attribué à l'édition par l'éditeur lui-même, soit séquentiellement soit selon sa propre organisation,
- Le dernier segment, nommé K correspond à une clé de vérification calculée via l'algorithme Modulo 11[190], de manière à limiter à 1% le risque des erreurs de frappe sur le reste de la chaine.

S'ajoute également à l'identifiant un principe de souplesse permettant une infinité de possibilités : étant donné que le nombre de chiffres par segment n'est pas figé, l'on pourra établir un quota de publications autorisées plus limité à un éditeur en lui attribuant un numéro plus long. Par exemple pour une petite association française qui n'aura pas plus de 100 publications en tout, on pourra lui attribuer un préfixe d'éditeur à 5 chiffres (ce qui logiquement en laisse 2 pour les numéros de publication : de 01 à 99).[191]

La syntaxe est donc représentée par une suite de chiffres, comme par exemple :

Exemple : ISBN 979 10 000001 1 5

L'espacement entre les segments, s'il est pratique pour interpréter humainement un ISBN, ne sera pas interprété par la plupart des systèmes. Lors de leur intégration en EAN, des tirets entre les segments sont ajoutés avant d'être exprimés en codes-barres.

ISBN-A

Il s'agit d'un ISBN actionnable (A pour actionnable), constituant une sorte de synthèse entre un ISBN et un identifiant DOI, ce qui lui permet d'être déréférencé dans le système DOI et donc de s'intégrer à cette base-là.[192] Il fait bien entendu partie des identifiants ISO.

La syntaxe ISBN-A est la suivante :

- Une succession de préfixes : le préfixe DOI « 10 » ; le préfixe ISBN (dans le cas d'un ISBN 10 : 978, dans le cas d'un l'ISBN13 : 979.) ; le préfixe de la zone géographique et de l'éditeur (attribué par l'Agence de registre ISBN) : de 2 à 8 chiffres,
- Le slash qui sépare les préfixes et suffixes,
- Le suffixe du numéro d'édition attribué par l'éditeur ainsi que la clé de vérification : de 2 à 6 chiffres puis le chiffre clé.

Ce qui donne un format comme dans l'exemple suivant :

Exemple : 10.978.0000001/018

---

[190] L'algorithme Modulo 11 est relativement simple à calculer : on va multiplier par 9 le premier chiffre du segment A, par 8 le deuxième, par 7 le troisième, etc. jusqu'à arriver au dernier chiffre avant la clé qui sera multiplié par 1. Ces chiffres seront additionnés et la clé de vérification sera le chiffre qu'il faudra lui rajouter pour que le total soit divisible par 11.

[191] HONORE, Suzanne. *Op. Cit.*

[192] Doi.org, *DOI System Factsheet* [en ligne] disponible sur https://www.doi.org/factsheets/ISBN-A.html [consulté le 02/05/2017]

ISSN

L'ISSN, *International Standard Serial Number* est l'équivalent de l'ISBN pour les publications continues. Lui aussi fait partie des identifiants ISO. Sont concernés les journaux, publications annuelles, revues, magazines, collections, etc. Il s'applique également aux sites web, aux bases de données, aux blogs, et dans d'autres cas particuliers. L'EAN l'intègre également même s'il doit le prolonger pour qu'il rentre dans le format des 13 chiffres : en effet, l'ISSN n'en possède que 8. Ceux-ci sont attribués de manière opaque sans lien avec le contenu ou la provenance.[193] L'attribution d'un ISSN est gratuite, mais à l'avenir certains centres pourront demander une contrepartie financière.

- Il commence par la mention ISSN,
- Il comprend 8 chiffres séparés en deux groupes de 4 par un tiret,
- Les 7 premiers chiffres sont opaques,
- Le dernier chiffre est une clé de contrôle générée par l'algorithme Modulo 11, et peut être un X lorsque le résultat de cet algorithme correspond à 10.

Exemple : 0029-4713

ISSN-L

ISSN-L pour ISSN de liaison est un ISSN qui permet de faire le lien entre les différents supports d'une même publication. Il s'agit ici d'un rassemblement des versions papier, numérique… c'est à dire, de différentes « *medium version* ». Il a été mis en place afin de faciliter la citation et la recherche autour d'une même revue.[194] L'ISSN propose déjà un lien entre les différentes représentations d'une même série (par exemple, la version numérique et la version papier), mais ce lien est explicité dans les métadonnées associées aux identifiants dans le registre alors qu'ici il s'agit d'identifier un tout.

En termes de syntaxe, il reprend la syntaxe de l'ISSN, excepté que l'on mentionne en préfixe « ISSN-L ».

Exemple : ISSN-L 0029-4714

ISTC

International Standard Text Code est un système d'identifiants issu de la norme ISO 21047 :2009 qui s'applique à tous les contenus qui concernent du texte : articles de journaux, livres audio, livres imprimés, ebook, etc. Les agences d'enregistrement ISTC sont au nombre de dix actuellement dans le monde, chapeautées par l'International ISTC Agency. Ici il s'agit non pas d'identifier le produit éditorial (comme un ISBN ou un ISSN), mais le « travail », la pièce d'écriture pouvant être intégrée dans un tout à plusieurs occasions.[195] Au niveau du

[193] issn.org, *Comprendre l'ISSN* [en ligne] disponible sur http://www.issn.org/fr/comprendre-lissn/quest-ce-que-lissn/ [consulté le 02/05/2017]

[194] Issn.org, *L'ISSN-L pour les publications à supports multiples* [en ligne] Disponible sur http://www.issn.org/fr/comprendre-lissn/regles-d-attribution/issn-de-liaison/ [consulté le 02/05/2017]

[195] Istc-international.org, *About the ISTC, Overview and Scope of the ISTC* [en ligne] Disponible sur http://www.istc-international.org/about.aspx [consulté le 02/05/2017]

modèle FRBR, l'ISTC se situe dans les entités du groupe 1 : l'expression.[196] L'idée est de séparer la ressource de son producteur afin que celle-ci soit identifiée de manière unique (et non en référence à un producteur) de manière à accroitre sa citabilité[197].

Composé de 16 caractères, la syntaxe d'un ISTC est la suivante[198] :

- Un élément d'enregistrement composé de trois caractères correspondant à l'agence en charge de l'enregistrement,
- L'année d'enregistrement (4 chiffres),
- 8 chiffres attribués à l'œuvre textuelle séquentiellement à partir du moment où le recueil de métadonnées le concernant est soumis, vérifié comme unique et valide,
- La clé de contrôle calculée à partir du MOD 16-3 (ISO/IEC 7064).

Exemple : ISTC 0A9 2001 12B4A109 9

Il est préférable d'attribuer un ISTC à un produit textuel avant même sa publication : ainsi, un livre pourra bénéficier d'un ISTC antérieurement à sa publication et à son attribution d'un ISBN, ce qui permettra d'avoir l'ISTC dans les métadonnées liées à l'ISBN. Il faudra néanmoins prendre en considération que malgré les éditions multiples, un ISTC restera le même pour tout contenu textuel identique.

ISMN

Acronyme de *International Standard Music Number*, il sert à identifier toutes les publications musicales imprimées effectuées mondialement, (c'est-à-dire les partitions) indépendamment de leur diffusion qu'elle soit payante ou gratuite.[199] Il est issu de la norme ISO 10957. Un même ouvrage peut donc posséder un ISBN et ISMN s'il publie des partitions dans son contenu. Ici il s'agit d'identifier une manifestation d'une œuvre, et non pas l'œuvre en elle-même.

En termes de syntaxe, il partage des éléments communs avec les types d'identifiants cités plus haut :

- Comme pour l'ISBN il est depuis 2008 constitué de 13 chiffres (avant 2008 de 10),
- Pour les publications antérieures à 2008, il était auparavant constitué d'un premier M puis de deux groupes de 4 chiffres correspondant au préfixe d'éditeur puis au numéro d'attribution (sans oublier un 9ème chiffre correspondant à la clé de vérification). Par exemple : M 2586 2569 6,
- Depuis 2008 et afin de faire correspondre le numéro EAN, les ISMN préexistants ont simplement été allongés en rajoutant 979- devant, et pour les nouveaux ISMN on édite toujours des numéros à 9 chiffres que l'ont fait précéder de 979-0.

---

[196] Plp. *FRBR : de l'expression bordel !* Site rda.abes.fr, 2013. [en ligne] Disponible sur https://rda.abes.fr/2013/08/30/frbr-de-lexpression-bordel/ [Consulté le 31/07/2017]

[197] Terme employé pour décrire les identifiants pérennes dans le document suivant : BERMES, Emmanuelle. *Des identifiants pérennes pour les ressources numériques, l'expérience de la BnF.* Bibliothèque nationale de France, 2006. 9p

[198] Istc-international.org, *About the ISTC, Basic structure and syntax* [en ligne] Disponible sur http://www.istc-international.org/about_structure_syntax.aspx [consulté le 02/05/2017]

[199] Ismn-international.org, *What is an ISMN?* [en ligne] Disponible sur http://www.ismn-international.org/whatis.html [consulté le 02/05/2017]

Exemple :

Avant 2008 : 979-M-2586-2569-6

Après 2008 : 979-0-2586-2569-6

ISWC

Un ISWC est un *International Standard Musical Work Code* issu de l'ISO 15707 :2001 qui identifie des œuvres musicales. Contrairement à l'ISMN[200], qui en identifie les publications papier, l'ISWC identifie vraiment le morceau en soi au sens de la composition (et non ses différents enregistrements comme l'ISRC, nous le verrons ci-après), afin de lier ses différents compositeurs/contributeurs etc. Il est géré par l'Agence Internationale ISWC. En France, l'agence d'attribution déléguée est la SACEM, et elle peut uniquement délivrer des ISWC à ses auteurs/contributeurs membres.

La syntaxe est la suivante :

- Le préfixe ISWC,
- La lettre T,
- L'identifiant de l'œuvre sur 9 chiffres (opaque),
- Un chiffre de contrôle.

La lecture est facilitée en ajoutant des tirets et des points, mais ceux-ci ne seront pas interprétés par la machine.

Exemple : ISWC T-034 698 521-1

ISRC

International Standard Recording Code correspond à un identifiant mis en place par l'International Federation of Phonographic Industry (IFPI) basé sur la norme ISO 3901 :2001, qui se focalise sur l'enregistrement en lui-même, fruit d'un regroupement (logique, artistique, thématique) de plusieurs personnes ou d'une seule à un moment donné pour produire un morceau. [201] Il s'agit donc d'identifier la manifestation d'une œuvre. Il concerne également les clips musicaux qui contiennent des éléments visuels animés. Ainsi, un même morceau (au sens composition) peut être lié à plusieurs ISRC en fonction du nombre de fois où il a été réinterprété, remixé, etc.[202] Il ne s'agit aucunement d'identifier un objet physique tel qu'un album ou un DVD, ou encore un enregistrement numérique dans un format en particulier.

Ce sont des agences d'attribution ISRC qui s'occupent de leur gestion auprès des producteurs dont ils sont les référents. Elles conservent pour chaque ISRC dans un registre et conformément à la norme ISO 3901 :2001 les informations suivantes sous forme de métadonnées : le code du pays du producteur et le code du producteur (qui sont cependant déjà indiqués dans le corps de l'identifiant même), les nom,

---

[200] Ismn-international.org, *4.ISWC (International Standard Musical Work Code)* [en ligne] http://www.ismn-international.org/publications/newsletter11/iswc.html [consulté le 02/05/2017]

[201] Scpp.fr, *Le code ISRC* [en ligne] Disponible sur http://www.scpp.fr/scpp/home/lascpp/guidepratiqueisrc/tabid/107/default.aspx [consulté le 02/05/2017]

[202] Blog.songtrust.com, *ISRC & ISWC – What's the difference?* [en ligne] http://blog.songtrust.com/songwriting-tips/isrc-iswc-whats-the-difference/ [consulté le 02/05/2017]

adresse et coordonnées du producteur, le département ou la personne à contacter si besoin, et enfin la date d'enregistrement de producteur dans la base.[203]

L'ISRC est construit au moyen de 12 caractères :

- 2 premiers caractères qui servent à identifier le pays du producteur,
- 3 caractères pour identifier le producteur lui-même,
- 2 chiffres correspondants aux deux derniers chiffres de l'année de dépôt,
- 5 chiffres correspondant à l'identification même du morceau.

Exemple : ISRC FR - Z03 – 98 - 00212

ISAN

ISAN veut dire *International Standard Audiovisual Number* et est un type d'identifiant ISO immatriculant les publications audiovisuelles. Les ISAN sont gérés par l'Agence internationale ISAN (ou ISAN-IA) qui regroupe et fédère de nombreuses organismes de production et de gestion de productions audiovisuelles.[204] Tous les ISAN sont accompagnés dans la base de données de métadonnées descriptives pouvant être plus ou moins complètes (titre, type d'œuvre, réalisateur, etc.) L'attribution d'un ISAN est payant pour chaque ressource identifiée. L'ISAN va être nommé différemment en fonction de son référent et de la granularité :

- L'ISAN-Œuvre identifie toute œuvre audiovisuelle quelle que soit sa nature,
- L'ISAN-Version ou V-ISAN quant à lui peut identifier une version précise d'une œuvre, avec un nombre illimité de V-ISAN attribués pour une même œuvre,
- L'ISAN-In dev peut être attribué au préalable à une œuvre en cours de création, et ce dès le début. Il devient par la suite un ISAN-Œuvre (ISAN-In dev activé) lorsque l'œuvre arrive à sa version définitive.

Cet identifiant est construit en 5 parties, chacune opaque, et contient en tout 24 caractères (sans compter le préfixe):

- L'acronyme ISAN précède le tout,
- Un numéro d'identification séquentiel à 12 caractères réunit en 4 blocs séparés par des tirets,
- Un ensemble de 4 chiffres correspondant au numéro d'épisode,
- Une lettre clé de contrôle,
- Un ensemble de 2 fois 4 chiffres correspondant à l'identification de la version,
- Une deuxième clé de contrôle.

Exemple : ISAN0000-3BAB-0000-0000-G-0000-0000-Q

[203] International ISRC Agency. *International Standard Recording Code (ISRC) Handbook.* IFPI Secretariat, 2009. Document PDF [en ligne] Disponible sur http://www.ifpi.org/content/library/isrc_handbook.pdf [consulté 02/05/2017]

[204] France-isan.org, *Le réseau international ISAN* [en ligne] Disponible sur http://www.france-isan.org/la-norme-isan/ [consulté le 02/05/2017]

DOI

DOI repose actuellement sur la norme ISO 26324, ayant été formalisée et ayant rejoint la multitude d'identifiants ISO en 2012.[205] Acronyme de *Digital Object Identifier* : c'est l'identifiant numérique d'un objet et non un identifiant pour un objet numérique.[206] L'International DOI Foundation (IDF) en est l'initiatrice et la gestionnaire de la fédération des agences d'enregistrement. Il s'agit d'identifier du contenu informationnel indépendamment de sa forme technique ou son mode d'accès afin de garantir une unicité de référence ainsi que son intégration dans un système d'information. Si à la base le DOI a été conçu pour identifier à la fois des contenus et des concepts (des documents, des institutions, des personnes, des évènements…) en pratique il est surtout utilisé actuellement pour les publications scientifiques (notamment avec le concours de son agence d'enregistrement affiliée Crossref)[207]. L'attribution d'un identifiant DOI favoriserait l'interopérabilité, la pérennité et l'accessibilité de données informationnelles grâce à leur référencement dans un système de gestion d'identifiant global. Si l'attribution est payante, un nombre croissant d'institutions (notamment des universités, et centres de recherche et développement, CNRS etc.) utilisent ce système pour identifier leurs ressources.

Les identifiants DOI sont opaques et fonctionnent grâce à deux codes distincts :

- D'une part le préfixe identifiant de manière unique une institution (un producteur de données). Celui-ci est suivi d'un slash qui sépare les deux codes,
- D'autre part le suffixe identifiant la ressource elle-même (attribuée par le producteur). Le producteur en est le seul responsable, il peut donc choisir d'attribuer n'importe quelle chaine de caractère à sa ressource, et éventuellement réutiliser des identifiants déjà implémentés dans son système.

Exemple : 10.1000/123456 ou encore 10.1000/issn.1234-5678

Le déréférencement d'un identifiant DOI s'effectue à l'origine via le système Handle qui ne prévoit une identification qu'en Utf-8. Actuellement, l'intégration d'un identifiant DOI dans une URL permet sa résolution par la formation d'une URI http correspondante via le résolveur en ligne (https://www.dx.doi.org/), mais il peut également être déréférencé directement via un *plug-in** Handle depuis un navigateur. La granularité n'est pas contrainte dans des spécifications particulières : chaque ressource identifiée peut être un tout ou une partie d'un tout. Il est du ressort des agences d'attribution de prévoir ou non des règles distinctes pour leur communauté, et surtout de faire en sorte que chaque référent possède des métadonnées assez complètes qui en précisent la nature et le type.

La German National Library of Science and Technology (TIB/UB) est la première agence d'enregistrement de DOI spécialisée sur la donnée scientifique et technique (STD-DOI) au monde, dans le cadre du projet Datacite.

---

[205] Doi.org, *DOI Handbook*, 1. Introduction [en ligne] Disponible sur https://www.doi.org/doi_handbook/1_Introduction.html [consulté le 02/05/2017]

[206] PASKIN, Norman, RUST, Godfrey. *Op. Cit.*

[207] Cela n'est cependant pas son seul domaine d'activité, il est également utilisé pour les jeux de données de Datacite, les documents audiovisuels avec EIDR, etc.

ISCI

L'identifiant International Normalisé des Collections s'applique en particulier aux biens culturels et est issu de la norme NF ISO 27730 publiée en 2012. Celle-ci a vocation à identifier de manière unique « *chaque collection, fonds et série archivistique et des parties de collections, de fonds et de séries archivistiques.* »[208] Elle ne propose cependant pas de méthode pour établir des relations entre les collections et sous-collections. L'ISCI fait écho avec l'ILII de la norme ISO/DIS 20247 (*International Library Item Identifier* ou Identifiant International pour élément de collection de bibliothèques), qui est encore au stade de projet, et qui correspond au même type de besoin.

L'identifiant ISCI se construit grâce à l'identifiant ISIL, auquel il est très lié, qui identifie les institutions culturelles telles que les bibliothèques et services d'archives. La syntaxe est la suivante :

- L'identifiant ISIL,
- Un séparateur (deux points),
- L'identifiant ISCI.

Exemple : FR-751041001:Casadesus1

ISLI

L'International Standard Link Identifier, à l'origine ISDL (*International Standard Document Link*) est un identifiant issu de la norme ISO 17316 servant en premier lieu à nommer les liens entre des entités identifiées par ailleurs qu'il s'agisse de ressources, concepts, organisations, documents web… L'ISLI est matérialisé par la triade Source-Cible-Lien qui relie les composants d'une relation.[209] Il identifie donc un lien unidirectionnel : d'une source à sa cible uniquement et non un aller-retour. Les identifiants sont attribués par l'agence d'enregistrement d'ISLI, ISLI-RA, qui s'occupe également de maintenir le registre, proposer un service de résolution, maintenir des listes de types de sources et de cibles, développer de la documentation, etc. Cette agence possède un serveur d'allocation des codes services (SCAS) et un système de gestion globale des données ISLI (GIDC). A l'origine, le standard ISLI dérive d'un besoin éditorial de pouvoir qualifier les liens existants au sein de produits de type crossmédia[210].

La syntaxe est composée de trois champs [211]:

- Le code service (6 caractères) : il est attribué par l'ISLI-RA,

[208] Boutique Afnor, NF ISO 27730 Août 2012. *Information et documentation – Identifiant international normalisé des collections (ISCI)* [en ligne] Disponible sur https://www.boutique.afnor.org/norme/nf-iso-27730/information-and-documentation-international-standard-collection-identifier-isci/article/701623/fa158868 [consulté le 22/05/2017]

[209] ISSN.org, *International Standard Link Identifier (ISLI).* 2014. [en ligne] Disponible sur http://www.issn.org/newsletter_issn/international-standard-link-identifier-isli/ [consulté le 22/05/2017]

[210] Le cross-média correspond à un produit éditorial mis en scène dans un univers spécifique dont la richesse peut s'apprécier à travers différents médias : un livre complété par une bande audio, un film à la suite d'une Bande Dessinée, etc. Un bon exemple de cross-média est l'ensemble des produits dérivés, productions éditoriales, films, sites, jeux et œuvres qui découlent des univers de Tolkien ou de J.K Rowling, qu'elles soient de sources officielles (l'auteur ou le scénariste) ou sous forme de fan-art.

[211] Isli-international.org. *ISLI Handbook*. [en ligne] Disponible sur http://www.isli-international.org/irap/web/navigation/toNavigationPage/13?selectedLanguage=en_US [consulté le 22/05/2017]

- Le lien du code (longueur variable) : il définit le lien qu'il y a entre la source et la cible. Il peut être divisé en deux, un premier lien qui concerne une œuvre dans son ensemble, un second qui peut correspondre à un fragment d'un œuvre,
- Le caractère de contrôle (1 caractère).

Exemple : ISLI 116063-45200862937914734264430 01-9

ISNI

L'ISNI est *l'International Standard Name Identifier*. Il est définit par la norme ISO 27729:2012. Il a pour vocation d'identifier les personnes et les organismes.[212] Nous entendons par personne une identité publique correspondant à des gens impliqués dans la production/diffusion d'un contenu informationnel, artistique, intellectuel. Il est à noter qu'à chaque identité est attribué un identifiant, même s'il s'agit in fine de la même personne (par exemple, quelqu'un qui utiliserait plusieurs pseudonymes pour exercer des activités différentes). L'ISNI est géré par l'Agence Internationale de l'ISNI (ISNI-IA), mais c'est l'Agence Internationale d'Attribution (ISNI-AA) qui s'occupe d'attribuer les identifiants concrètement. Cette dernière comporte des agences d'enregistrement (dont la BnF) ainsi que des membres simples.

La construction s'est faite sur la base du VIAF, avec qui un échange régulier permet de mettre à jour et comparer les données recueillies. Le VIAF fournit une « super notice d'autorité » pour une entité donné et fait le lien avec les différentes notices nationales et régionales tandis que l'ISNI procure un identifiant pérenne et unique pour la désigner.

La syntaxe de l'ISNI comporte 4 lots de 4 chiffres chacun (16 chiffres en tout), incrémentés de manière séquentielle pour une identification opaque. Le déréférencement est possible via un champ de recherche dans l'URL http://isni.oclc.nl/ qui propose également une recherche par nom directement.

Exemple [213] :

ISNI:	0000 0001 2143 0518
Name:	Massimiliano I, Holy Roman Emperor Maximilian 01, Holy Roman Emperor Maximilian I, Holy Roman Emperor (Roman emperor, 1459-1519) Maximiliano I, Holy Roman Emperor Maximilien I, Holy Roman Emperor

ISIL

L'International Standard Identifier for Libraries and Related Organizations est issu de la norme ISO/DIS 15511. Il s'agit d'identifier des organismes, en l'occurrence des bibliothèques et des services d'archives, afin de permettre une

[212] Bnf.fr, *ISNI (International Standard Name Identifier)* [en ligne] Disponible sur http://www.bnf.fr/fr/professionnels/isni_informer.html [consulté le 02/05/2017]

[213] Cet exemple est tiré de la recherche suivante sur isni.oclc.nl : http://isni.oclc.nl/DB=1.2/SET=3/TTL=1/PRS=DEFAULT/NXT

interopérabilité plus générale dans les échanges de données entre les différentes structures à travers le monde[214]. Une fois enregistrés dans la base, les identifiants des structures peuvent être associées à des métadonnées descriptives (notamment ISDIAH pour les services d'archives, qui permet d'ajouter des informations pratiques sur le service : collections, historique du service, liens vers des sites web, etc.). Ces identifiants peuvent également être utilisés pour identifier un dépôt d'archives dans une description EAD.

La syntaxe change en fonction des pays selon une norme qui leur est propre, qui sont notifiés au début de l'identifiant avec un code à deux chiffres (pour la France, FR, Italie, IT, Belgique, BE, etc.).

En ce qui concerne la France, par exemple, le code ISIL sera constitué du numéro du RCR (Répertoire des Centres de Ressources), précédé de FR-.

UUID

Les *Universally Unique Identifiers* sont des séries de caractères qui peuvent être générées à l'envie et dont l'unicité est d'une probabilité très élevée à chaque fois. Ici, il n'est point question d'autorité nommante ou de métadonnées, les UUID sont basées sur la norme ISO/IEC 9834-8 :2008 et sont générées par algorithme. Ces identifiants sont totalement opaques et leur utilisation est simple, accessible à tous et entièrement gratuite. Ils sont très utilisés dans la programmation et le développement informatique où ils se révèlent très intéressants. Le site UUID Generator.net propose un nouvel identifiant à chaque réactualisation de la page.[215]

On peut voir qu'il y a différentes versions de l'UUID : la version 1 génère un nombre en se basant sur l'adresse MAC de l'ordinateur qui fait la requête. La version 4 de l'UUID est quant à elle un nombre généré au hasard sans lien avec le contexte de requête. Il n'y a cependant aucune garantie que le nombre soit réellement unique : en effet, si les probabilités estiment à un pourcentage infinitésimal le risque de tomber deux fois sur le même nombre, il n'y a pas de risque zéro. Ils ne sont donc pas d'une stabilité et d'une fiabilité sans faille pour ce qui est de l'identification pérenne d'un contenu.

La syntaxe se compose de 32 caractères délimités en 5 blocs de respectivement 8 chiffres, 3 blocs de 4 chiffres et 12 chiffres.

Exemple d'un UUID version 4 : df0ad42d-fbeb-44ea-81a7-6c14b83f1e53

2.2.2. Les identifiants globaux issus d'initiatives individuelles

ARK

L'*Archival Resource Key* est un identifiant de la California Digital Library (CDL) qui peut également s'appliquer à tous types de contenus, des objets physiques, des concepts, des documents web, des personnes ou institutions, etc. Il nécessite systématiquement l'utilisation du protocole DNS, et est actionnable

[214] Isil.arch.be, *Isil* [en ligne] Disponible sur http://isil.arch.be/?changelang=1&lang=fr&view=register&iid= [consulté 02/05/2017]

[215] Voir https://www.uuidgenerator.net/

uniquement sous la forme d'une URL dans le nom de domaine de l'institution donnant accès aux ressources.

On distingue deux acteurs identifiés qui prennent part à la mise en place d'identifiants ARK :[216]

- La *Name Mapping Authority* (NMA) est l'institution qui donne accès aux ressources. Elle peut être différente de celle qui attribue les identifiants. Nous partons ici du principe que si l'organisation qui donne accès aux données peut évoluer, les ressources, elles, n'évoluent pas. Cette autorité d'adressage (*Name Mapping Authority Host*, NMAH) va donc apporter un nom de domaine qui va permettre d'accéder à l'objet identifié ou sa représentation, mais celui-ci sera remplaçable.[217] Par exemple, Gallica ayant été désigné comme le NMA, l'URL de ses ressources identifiées en ARK commencera de la manière suivante : http://gallica.bnf.fr/ark:/.....
- Le *Name Assigning Authority* (NAA) est l'institution attribuant les identifiants ARK. Chaque institution attributaire obtient un NAAN (*Name Assigning Authority Number*) qui l'identifie de manière unique et pérenne, et ce numéro fait partie intégrante de l'identifiant. Dans l'exemple précédent, le NAA et le NMA sont une seule et même institution. Cependant, si Gallica venait à fusionner avec une autre base et/ou changer de nom, le nom de domaine pourrait évoluer et cela n'aurait pas d'incidence sur la pérennité de l'identifiant ARK.

La syntaxe de l'identifiant ARK est la suivante :

- Le préfixe « ARK »,
- Le NAAN (*Name Assigning Authority Number*) qui correspond au code d'identification de l'institution attribuant le numéro,
- Le nom ARK attribué à la ressource par l'autorité nommante. Celui-ci peut être très long, et être composé lui-même de différentes sous parties afin d'intégrer des systèmes d'identification déjà existants. Il ne doit pas contenir de voyelles, mais par contre il peut contenir une clé de contrôle à la fin,
- L'ARK peut être complété par des qualifieurs, ou qualifiants, c'est-à-dire des informations supplémentaires à l'identification même servant au cours du déréférencement, qui sont interprétés par le résolveur de l'autorité d'adressage. Ces qualifieurs sont spécifiques à chaque NMAH. Ils peuvent concerner la granularité (par exemple, permettre l'accès à une partie seulement du contenu) ou la réalisation d'un service (par exemple, l'accès à une version du document en particulier). Cela permet notamment d'avoir une identification majeure pour les contenus globaux et ensuite gérer la granularité et le service sans avoir à attribuer de nouveaux identifiants à chaque fois.

Exemple : ark 12148/0000000000000/version1

---

[216] Wiki.ucop.edu, *ARK (Archival Resource Key) Identifiers* [en ligne] Disponible sur https://wiki.ucop.edu/display/Curation/ARK [consulté le 02/05/2017]

[217] Bnf.fr, *ARK (Archival Resource Key)* [en ligne] Disponible sur http://www.bnf.fr/fr/professionnels/issn_isbn_autres_numeros/a.ark.html [consulté le 02/05/2017]

SICI

Il s'agit du *Serial Item and Contribution Identifier*, dérivé du standard ANSI/NISO Z39.56. Celui-ci est *deprecated*, c'est-à-dire qu'il n'est plus utilisé et que le standard originel de la NISO (*National Information Standard Organisation*) a été supprimé. Il est cependant intéressant de le mentionner au vu de son système d'identification intégrant et réutilisant d'autres identifiants.

Le SICI était une extension de l'ISSN. Il servait à identifier des fragments d'un contenu.[218] Mis en place par le Serials Industry Systems Advisory Committee (SISAC) en 1991, il était de longueur variable et s'associait avec d'autres types d'identifiants comme le PII, le DOI ou encore l'URN. Dans le cas de l'ISSN et du DOI, le SICI pouvait directement intégrer l'identifiant en son sein. Pour ce qui est de l'URN, l'INFO ou le PII, le SICI se voyait attribuer un préfixe.

Le SICI était structuré généralement en 3 parties[219] :

- Une partie « objet » qui réutilisait l'identifiant de l'entité duquel le contenu à identifier était extrait. Par exemple, un ISSN ou un DOI,
- Une partie « contribution » entre chevrons (<>) qui délimitait la partie que l'on souhaitait identifier exactement. Lorsque le SICI était transposé en INFO, URN ou autre les chevrons étaient changés en signe pourcentage (%),
- Une partie de « contrôle » qui comportait le CSI (*Code Structure Identifier*), dont le rôle était de vérifier la validité de la structure de cet identifiant ; le DPI (*Derivative Part Identifier*), qui lui vérifiait la validité du fragment concerné ; le MFI (*Media Format Identifier*), un identifiant de deux caractères qui donnait le support de présentation du contenu ; le numéro de version ; et enfin la clé de contrôle générale de l'identifiant.

Exemple avec un numéro d'ISSN [220]:

0095-4403(199502/03)21:3<12:WATIIB>2.0.TX;2-J

Exemple avec un DOI :

10.1002/0002-8231(199601)47:1<23:TDOMII>2.0.TX;2-2

Exemple avec un INFO :

info:sici/1046-8188(199501)13:1%3C69:FTTHBI%3E2.0.TX;2-4

Si le SICI était très utile pour les publications en série, il se limitait à elles. De plus, il n'était pas utilisable si la ressource à identifier n'était pas encore assignée à une publication en particulier.

BICI

Le *Book Item and Component Identifier* est la version du SICI pour les livres. Il est censé contourner l'écueil évoqué plus haut du SICI concernant la limitation des publications en série. Créé par la Book Industry Communication (avec le soutien

[218] GREEN, Brian, BIDE, Mark. *Op. Cit.*

[219] BLIXRUD, Julia. *SICI and BICI : Identifiers for Serials and Books*. Association of Research Libraries and Chair, NISO Subcommittee AP. [en ligne] Disponible sur: https://www.cendi.gov/presentations/ref_link_blixrud.ppt *[consulté le 05/05/2017]*

[220] Exemples tirés de la page Wikipédia sur le SICI [en ligne] Disponible sur https://en.wikipedia.org/wiki/Serial_Item_and_Contribution_Identifier [consulté le 05/05/2017]

du fonds de recherche BNB de la British Library) il permet d'identifier des parties d'un livre (extraits, chapitre, paragraphes) qu'il soit physiquement contenu dans le livre ou indépendant : des illustrations, cartes, préface, index, bibliographie, tableaux, ou tout autre contenu textuel ou non-textuel. Il permet également d'identifier des parties d'ouvrage qui ne comportent pas de chapitre comme par exemple les encyclopédies. Par contre il ne concerne pas la littérature grise et les documents d'activités tels que les rapports techniques : il ne s'agit que de publications.

Dans la syntaxe, il reprend plus ou moins le SICI avec trois parties (objet, contribution, contrôle) en l'utilisant à la place de l'ISBN. Le segment de contrôle contient cependant un *Component Type Identifier* (CTI) et un *Standard Version Number* (SVN) qui lui sont ajoutés.[221] Le CTI est un chiffre de 0 à 5 indiquant le type d'extrait : 0=ensemble, 1=élément de couverture, 2= subdivision de corps du texte chapitre, section, acte, etc., 3= objet discret type figure, tableau, etc., 4= éléments de fin type index, bibliographies, etc. 5=élément supplémentaire.

Exemple : BICI: 0521416205(1993)(10;EAAWL;234-261)2.2.TX;1-1

Si un « *draft* » (brouillon, première version) a été mis en place en 2000 par NISO, il semblerait que la version définitive n'ait pas encore abouti.

ELI

ELI pour *European Legislation Identifier* est un identifiant qui concerne uniquement les textes de loi Européens.[222] Sa vocation est entièrement centrée sur un objectif de meilleure diffusion des textes de loi aux citoyens et aux organisations. Il s'agit d'une identification non obligatoire pouvant être attribuée parallèlement aux systèmes d'identification spécifiques à chaque institution. L'ELI permet en outre un échange de données liées aux lois facilité entre des systèmes hétérogènes. La France l'a implémenté depuis 2002 sur les Journaux Officiels. Toutes les Mesures Nationales d'Exécution antérieures à cette année se sont vues attribuer des ELI rétroactivement en 2015. Légifrance l'a quant à lui intégré dans son fonctionnement en 2014. Il existe deux types d'ELI, l'ELI de diffusion et l'ELI de production qui utilise le NOR (alias unique permettant la rétrocompatibilité avec les anciens systèmes).

Chaque pays a une syntaxe adaptée. La syntaxe pour la France est organisée de la manière suivante[223] :

- Le préfixe /eli/,
- Le type de loi (arrêté, circulaire…),
- La date Année/Mois/jour (pas de 0 pour les mois et les jours de 1 à 9),
- L'identifiant naturel : c'est là que l'on constate la divergence entre ELI de diffusion et ELI de production, dans l'ELI de diffusion l'identifiant naturel correspond à l'année (aaaa) tiret le numéro du texte dans l'année, dans l'ELI de production il s'agit tout simplement du NOR (12 caractères),

[221] BLIXRUD, Julia. *Op. Cit.*

[222] Eli.fr [en ligne] Disponible sur http://www.eli.fr/fr/index.html [consulté le 22/05/2017]

[223] Eli.fr, *Eléments constitutifs des URI pour la France*. [en ligne] Disponible sur http://www.eli.fr/fr/constructionURI.html [consulté le 22/05/2017]

- Le domaine (Nom normalisé, servant à remplacer l'identifiant naturel lorsque celui-ci n'est pas disponible),
- La version : soit jo (version originale), soit lc (version consolidée d'un texte législatif),
- Le niveau : la granularité de l'identification (texte complet ou article dans le texte),
- La date de la version : paramètre facultatif, il s'agit de fixer une version particulière du texte. Si ce critère n'est pas précisé c'est la version actuellement en vigueur qui est retournée,
- La langue : fr (plusieurs possibilité suivant les pays qui possèdent plusieurs langues officielles). Ce critère est toujours composé de deux caractères,
- Le format (pdf, html…).

Tous ces critères sont séparés par des slashs.

Exemple : eli/decret/2017/8/25/INTD1713523D/jo/texte

Handle (HDL)

Si le système Handle est à la base une spécification technique permettant de déréférencer des identifiants (notamment les identifiants DOI), il possède lui aussi ses propres identifiants que l'on distingue par le préfixe HDL.

Il est géré par la DONA Foundation de manière globale (elle est gestionnaire du Global Handle Registery, GHR). Cependant les préfixes des identifiants sont attribués par la Corporation for National Research Initiatives (CNRI) en son nom.[224] Cette attribution est payante annuellement.

La syntaxe Handle [225] correspond tout à fait à celle utilisée pour les DOI : un préfixe et un suffixe séparés par un slash central.

Exemple : HDL : 12345/4561

L'URL http://hdl.handle.net/ permet de résoudre les identifiants Handle (on mettra directement après le slash l'identifiant sans le préfixe HDL). Cette URL renvoie l'usager via une table de correspondance à l'adresse URL qui contient la ressource demandée. La DONA Foundation possède un serveur central pour le Registre Handle Global (RHG) et de nombreux serveurs locaux.

L'avantage du système réside essentiellement dans le fait qu'il sépare localisation et identification, comme chez les identifiants DOI, de manière à permettre une souplesse d'évolution propice à la pérennité des références.

Dans une page explicative de leur site, les responsables DOI expliquent en quoi DOI va plus loin que ce qui est proposé par les identifiants Handle :[226] il apporte une rigueur de sélection qui accroit de fait l'interopérabilité et la pérennité des identifiants, en plus d'un management, d'un support technique et d'une gestion des métadonnées plus contraignante.

---

[224] Handle.net, *HDL.NET® Information Services* [en ligne] http://www.handle.net/index.html [consulté le 02/05/2017]

[225] ARCHIMBAUD, Jean-Luc. *Op. Cit.*

[226] Doi.org, *Factsheet, DOI® System and the Handle System®* [en ligne] Disponible sur http://www.doi.org/factsheets/DOIHandle.html [consulté le 02/05/2017]

ORCID

Récemment relié à l'ISNI, ORCID est une identification spécifique pour les chercheurs (auteurs et contributeurs de l'enseignement supérieur) qui prend une grande ampleur dans le domaine. Il réunit notamment les gros acteurs d'édition scientifique tels que Elsevier, Nature Publishing, Wiley et Springer. Chaque chercheur possède une identification propre via un enregistrement rapide auprès d'ORCID. L'identifiant est apposé sur toutes les étapes de *versioning* constituant le déroulement de la publication scientifique (manuscrit, soumission des écrits, publication, etc.) et lie de manière permanente et continue l'auteur et son travail.[227] Cela permet donc de garantir la reconnaissance des travaux à leurs auteurs tout au long du processus, avant même la publication. La collaboration avec l'ISNI garantit à l'ORCID l'utilisation d'une partie de ses espaces de numéros afin qu'il puisse identifier de manière unique et pérenne les personnes tout en permettant l'interopérabilité des deux systèmes. L'ORCID s'est adapté pour conformer ses identifiants au standard de la norme ISO 27729:2012.

La méthode d'identification est assez similaire à celle des DOI. L'ORCID est composé de 16 chiffres, attribués de manière aléatoire par le registre ORCID de manière à ce qu'ils n'entrent pas en conflit avec les numéros de l'ISNI. Le dernier caractère est systématiquement un caractère de contrôle selon le système du Modulo 11, tel que nous l'avons vu précédemment.

Exemple : ORCID 0000-0001-9873-1538

L'identifiant est totalement opaque, et ne comporte pas d'information sur le chercheur. Ainsi, aucune information liée à des éléments de carrière ou d'orientation professionnelle du référent n'est intégrée, ce qui permet également la conservation de l'identifiant tout au long de la vie professionnelle de la personne identifiée. L'idée est aussi de conserver une certaine confidentialité dans le cas où l'anonymat des chercheurs devrait être respecté.[228]

L'ORCID est transformable en URI http lorsqu'on lui ajoute le préfixe http://orcid.org/. L'identifiant qui suit sera également découpé par des tirets tous les 4 chiffres afin d'améliorer la lisibilité de l'ensemble. Il y a donc tout de même une volonté chez l'ORCID de rendre l'identifiant opaque manipulable plus aisément par des humains.

IPI

Mis en place par la Confédération Internationale des Sociétés d'Auteurs et Compositeurs (CISAC), *l'Interested Party Information* est un identifiant de personne lié à la gestion des droits moraux et patrimoniaux d'une œuvre musicale. IP est plus souvent utilisé pour désigner l'identifiant lui-même. La base est maintenue et administrée par la société d'auteurs suisse SUISA qui avait développé la version initiale, la CAE (Compositeur Auteur, Editeur), qui contenait 9 chiffres.[229]

---

[227] Orcid.org. *Distinguish yourself in three easy steps*. [en ligne] Disponible sur https://orcid.org/ [consulté le 31/07/2017]

[228] Orcid.org. *Structure of the ORCID Identifier.* [en ligne] Disponible sur https://support.orcid.org/knowledgebase/articles/116780 [consulté le 31/07/2017]

[229] Fr.cisac.org. *IPI*. [en ligne] Disponible sur http://fr.cisac.org/Nos-Activites/Services-d-information/IPI [consulté le 06/06/2017]

Une œuvre ne peut pas avoir de code ISWC sans que ses contributeurs ne soient identifiés via le code IPI.[230] C'est un identifiant de nom, pas de personne, donc chaque pseudonyme a son IPI : ils sont cependant tous connectés à un identifiant de personne. Il y a néanmoins une gestion de la confidentialité : le lien entre les différents IPI est soumis à des droits d'accès. Par contre l'identifiant gère très mal les anciens créateurs pour lesquels les droits ne s'appliquent pas : par exemple ceux dont les œuvres qui sont tombées dans le domaine public, car il a pour vocation première de permettre le paiement des royalties et des droits patrimoniaux.

La syntaxe est composée de trois segments :

- Un en-tête (1 seule lettre),
- Un numéro d'identification (9 chiffres),
- Une clé de contrôle (1 chiffre).

Exemple : L-000000000-0

L'IPI fait néanmoins redondance avec l'ISNI dans certains domaines, puisqu'ils identifient tous deux des personnes du secteur culturel (entre autres). L'ISNI couvre déjà de nombreux secteurs, pourtant il s'étend progressivement sur celui de la musique qui est pourtant le domaine privilégié de l'IPI. Si l'IPI a une seule méthode d'enregistrement auprès d'une agence de référence, n'importe quelle organisation peut candidater à l'attribution d'ISNI pour identifier ses auteurs/compositeurs : ils ont une base globale à OCLC. Par contre ils se rejoignent sur l'approche de la gestion des liens entre pseudonymes et identité réelle. A terme, l'IPI devrait se rallier à l'ISNI afin de s'accorder sur une identification universelle d'agents.[231]

LSID

Les *Life Science Identifiers* (identifiants des sciences de la vie) ont été mis en place en 2003 par un Consortium aujourd'hui séparé, le *Interoperable Informatics Infrastructure Consortium* (I3C) comprenant plusieurs grands acteurs tels que IBM, Oracle, Sun Microsystems et l'Institut de Technologie du Massachussets. Le principe était de permettre l'échange de données scientifiques entre les universités, les chercheurs et les laboratoires privés afin de faciliter les efforts collaboratifs sur des projets d'avancées de la science, notamment pharmaceutique et médicale. Un système de résolution de ces identifiants a été mis en place, le LSRS (*LSID Resolution System*) qui retourne les métadonnées en RDF, permettant la réintégration de celles-ci dans d'autres bases.[232]

Ces identifiants ont vocation à nommer (et non à localiser) des entités constitutives du monde biologique et sont enregistrées dans la base des URN.

La syntaxe est la suivante[233] :

---

[230] GREEN, Brian, BIDE, Mark. *Op. Cit.*

[231] Blog.tagyourmusic.com. *Musimorphoses – Vers l'atteinte d'un standard d'indexation universel et ouvert*. [en ligne] Disponible sur http://blog.tagyourmusic.com/fr/musicmetadata/ [consulté le 06/06/2017]

[232] GARRITY, George M., THOMPSON Lorraine M., USSERY, Dave W., PASKIN, Norman, BAKER, Dwight, DESMETH, Philippe, SCHINDEL, David E., ONG, Perry S. 7th meeting. *Study on the identification, tracking and monitoring of genetic resources.* UNEP/CBD/WG-ABS/7/INF/2, 2009. 98 p. [en ligne] Disponible sur https://www.cbd.int/doc/meetings/abs/abswg-08/information/abswg-08-abswg-07-inf-02-en.pdf [consulté le 29/11/2016]

[233] *Ibid.*

Urn : lsid : (identifiant de l'autorité) : (espace de nom) : (identifiant de l'objet) [: (version)]

Exemple : urn :lsid :ebi.ac.uk :SWISSPROT.accession :P34355 :3.

Le système LSID n'est pas payant, et est utilisé dans des bases spécialisées telles que Swiss-Prot, PubMed, GeneOntology. Les institutions qui ne souhaitent pas s'enregistrer mais qui veulent tout de même enregistrer des espaces de noms peuvent le faire grâce à la GBIF (*Global Biodiversity Information Facility*) qui prend les initiatives individuelle sous son identifiant d'autorité[234].

Les LSID ne sont pourtant très utilisés parce que des acteurs majeurs, tels que le NCBI (*National Center for Biotechnology Information*), ne souhaitaient pas s'y joindre car ils possédaient déjà des méthodes d'identification propres[235].

EIDR

L'Entertainment Identifier Registry Association (EIDRA) est une association d'acteurs de l'industrie de la production audiovisuelle gérant un identifiant qui lui est propre afin de coordonner l'identification des ressources audiovisuelles de manière globale et à moindre coût. Il est lié au DOI c'est-à-dire que toutes les ressources identifiées via cet identifiant sont également intégrées dans la base DOI.[236]

La syntaxe utilisée pour cet identifiant est la suivante :

- Un préfixe standard pour le registre EIDR qui est le 10.5240 (préfixe DOI),
- Un slash de séparation (propre au DOI),
- Une série de caractères alphanumérique (5 fois 4 chiffres, dont chaque ensemble est séparé des autres par un tiret),
- Un caractère de contrôle.

Exemple : 10.5240/0000-0000-0000-0000-0000-C

S'il fait redondance avec l'ISAN, un accord entre les différents acteurs a été trouvé [237]: ils offrent un service combiné entre les atouts d'ISAN (qui sont le réseau des agences d'enregistrement et leur service personnalisé), et le système d'identifiants d'EIDR qui dispose de tous les prérequis pour l'intégration dans la distribution numérique globale. Un même demandeur pourra alors faire une demande et obtenir les deux identifiants (il utilisera celui qui est le plus approprié à son besoin) et les deux seront liés dans le schéma du registre. L'EIDRA et l'ISAN-IA développent également des groupes de travail sur des problématiques partagées comme le développement des schémas de données ou la gestion des métadonnées.[238]

---

[234] *Ibid.*

[235] *Ibid.*

[236] Eidr.org, *About EIDR* [en ligne] Disponible sur http://eidr.org/about-us/ [consulté le 06/06/2017]

[237] Adami, BFI, CEPI, CineRegio, EIDR, et al. *Declaration on audiovisual work identifiers.* [en ligne] Disponible sur https://ec.europa.eu/licences-for-europe-dialogue/sites/licences-for-europe-dialogue/files/9-AV-identification.pdf [consulté le 06/06/2017]

[238] Eidr.org, *EIDR and ISAN to provide seamless registration of content IDs, agreement will leverage the strengths of both systems* [en ligne] Disponible sur http://eidr.org/eidr-and-isan-to-provide-seamless-registration-of-content-ids/ [consulté le 06/06/2017]

2.2.3. Les identifiants locaux

PII

Publisher Item Identifier est un identifiant créé par un regroupement de plusieurs acteurs majeurs de l'édition scientifique (Elsevier Science, American Physical Society, IEEE, etc.) et se fait appeler STI Group (*Scientific and Technical Information publishers*). Initialement il s'agissait de développer un outil d'échange interne entre les différents éditeurs membres, mais le PII a été repris par d'autres éditeurs externes qui le trouvaient intéressant pour leur propre usage. L'idée était d'identifier les articles avant leur publication afin de pouvoir les gérer avant même qu'ils ne soient assignés à une publication. Cela permettait notamment de répondre au problème posé par le SICI qui ne pouvait être attribué si l'extrait n'était pas déjà assigné à un ISSN. Le PII dérive du *Elsevier Standard Document Identifier* et du nombre ADONIS qu'il a d'ailleurs remplacé. Il n'est pas géré uniformément et il n'y a pas de registre central. [239]

La syntaxe, se voulant opaque, se compose de 17 caractères :

- Un chiffre correspondant au type de source de publication,
- Le code ISSN ou ISBN,
- Dans le cas d'éléments sériels, les deux derniers chiffres de l'année d'attribution de l'identifiant,
- Un nombre unique pour la ressource identifiée,
- Un caractère de contrôle.

Exemple : S0960-9822(11)01319-4

Les éditeurs doivent jongler entre SICI et PII car le PII ne permet pas de retrouver les articles publiés, et n'est pas rétroactif sur les publications antérieures. Il est simplement fait pour garder une trace du cycle de vie de la ressource et n'a pas d'affordance.

IMDB

L'Internet Movie Database (IMDb) est, comme son nom l'indique, une base de données collaborative en ligne qui référencie les productions audiovisuelles au niveau international. Lancée en 1990 par Col Needham, elle est rachetée en 1998 par Amazon et est encore très consultée. Son principe d'ouverture lui permet de saisir les opportunités d'être enrichie par les utilisateurs après vérification[240].

Les identifiants utilisés par l'IMDB servent pour deux types de référents : les titres de films, et les contributeurs de l'audiovisuel (acteurs, producteurs, réalisateurs…). Ils sont souvent « alignés » avec d'autres identifiants comme l'ISAN, l'EIDR ou encore l'IVA (les identifiants propres au site de *l'Internet Video Archive*), mais aussi des identifiants propriétaires tels que ceux d'Amazon, Sony, BFI, etc.

---

[239] GREEN, Brian, BIDE, Mark. *Op. Cit.*

[240] Page Wikipédia Internet Movie Database. [en ligne] Disponible sur https://fr.wikipedia.org/wiki/Internet_Movie_Database [consulté le 06/06/2017]

L'identifiant IMDB est composé d'un acronyme signifiant le type de référent suivi de 6 chiffres séquentiellement attribués[241] :

- « tt » pour les entreprises (title),
- « nm » pour les personnes (name),
- « co » pour les organisations/entreprises (company),
- « ch » pour les classements (charts),
- « ev » pour les évènements (events).

Sous forme d'URI http, cet identifiant est précédé du nom de domaine ainsi que de l'espace de noms correspondant au type.

Exemple : identifiant local : /tt123456/, identifiant URI http : www.imdb.com/title/tt123456/

TAG URI

Les TAG URI sont des identifiants « maison » qui sont utilisés pour [242]:

- désambiguïser des ressources,
- garder un niveau d'affordance permettant à des humains de les manipuler,
- être indépendant d'un système de résolution d'identifiants,
- être indépendant d'un enregistrement à un registre central,
- et enfin de réduire les frais d'attribution.

C'est une alternative aux UUID qui sont trop opaques, aux DOI qui sont trop chers, aux http URL qui impliquent la localisation pointée de la ressource (pas évident pour des concepts immatériels) et qui de plus sont moyennement pérennes, et enfin aux PURL qui nécessitent d'être dépendants de OCLC.

Comme l'explique de manière très ludique dans le tutoriel disponible sur le site Tag uri.org[243], la syntaxe de base de l'identifiant est créée à la volée par l'utilisateur qui va choisir lui-même des valeurs permettant d'identifier de manière unique :

- Le préfixe « tag : »,
- Un identifiant pour l'utilisateur (dans l'exemple : une adresse email), si cette information seule ne permet pas de garantir l'unicité, l'entièreté de l'identifiant sera quand même unique grâce aux autres éléments contenus,
- Une date que l'on mettra au format préconisé par l'ISO 8601 (2017-05-05) ou bien 2017-05 s'il s'agit du premier jour du mois ou 2017 s'il s'agit du premier jour de l'année. Le choix de la date peut éventuellement être en rapport avec la date d'attribution,
- Un nom unique pour la ressource ou l'objet : il peut être signifiant ou opaque, à la convenance de l'utilisateur.

Chaque partie de l'identifiant sera séparé des autres par une virgule.

---

[241] Page Wikidata de IMDb ID (P345) [en ligne] Disponible sur https://www.wikidata.org/wiki/Property:P345 [consulté le 06/06/2017]

[242] Tools.ietf.org, *The 'tag' URI Scheme.* [en ligne] Disponible sur https://tools.ietf.org/html/rfc4151#page-4 [consulté le 06/06/2017]

[243] Taguri.org, *Tag URI* [en ligne] Disponible sur http://www.taguri.org/ [consulté le 06/06/2017]

Exemple : tag :maurice.dupond@aol.fr,2017-05-05,00000001

La préconisation s'arrête là, c'est-à-dire que c'est aux utilisateurs de l'adapter à leurs besoins et aux objectifs de son utilisation.

OAI

Les identifiants OAI (*Open Archive Identifier*) sont des identifiants de l'Open Archive Initiative « surnuméraires », c'est-à-dire qui viennent se superposer à d'autres identifiants (plus ou moins pérennes) des ressources, afin qu'elles soient uniformisées lors de leur intégration au sein d'un système OAI-PMH (*Open Archive Initiative - Protocol for Metadata Harvesting*). Ils font partie des espaces de nom URN (que l'on développe ci-après). Ils servent notamment à garantir une certaine unicité et pérennité des lots de métadonnées des ressources dans ces systèmes. L'Open Archive Initiative tient registre des fournisseurs de données, et en pratique tout le monde peut décider d'ouvrir un entrepôt OAI-PMH en s'enregistrant en tant que fournisseur.

La syntaxe d'un identifiant OAI ressemble à celle des identifiants DOI : elle reprend le *scheme* OAI puis ajoute l'identifiant du fournisseur de données (identifiant reçu lors de la déclaration de ce fournisseur), puis un identifiant local attribué par le fournisseur.[244]

Exemple : oai : arXiv.org : 0000000

PPN

Les identifiants PPN, pour *Pica Production Number* sont des identifiants locaux liés aux bases de données de la solution PICA, développée par l'OCLC. C'est notamment le cas de leur base bibliographique Worldcat. En France, l'ABES (Agence Bibliographique de l'Enseignement Supérieur) est notamment l'un des principaux utilisateurs, avec son système Sudoc (Système Universitaire de Documentation). Ces identifiants ne sont pas déréférençables tels quels mais peuvent l'être via les URL des structures qui l'utilisent par exemple http://www.sudoc.fr/PPN. Dans le Sudoc, un identifiant PPN peut également se référer à un auteur. Ainsi, le service IdRef fait le lien entre l'auteur et ses publications[245].

La syntaxe est composée de 9 caractères [246]:

- 8 caractères correspondant au numéro séquentiel de la notice dans la base,
- 1 chiffre de contrôle, ou bien un X.

Exemple : PPN 142914614

HALid et idHAL

HAL est un site de diffusion d'archive ouvertes 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,*

[244] ARCHIMBAUD, Jean-Luc. *Op. Cit.*

[245] ARCHIMBAULT, Jean-Luc. *Op. Cit*.

[246] *PPN*. Bibliopedia.fr [en ligne] https://www.bibliopedia.fr/wiki/PPN [consulté le 24/08/2017]

des laboratoires publics ou privés. »[247] C'est une plateforme mutualisée, projet issu de l'initiative du Centre pour la Communication Scientifique Directe (CCSD), bientôt rejoint par plusieurs établissements (INRIA, CEA, Institut Pasteur, CNRS, etc.) qui s'accordent sur la mise en place d'un service pour la recherche. HAL est connectée avec arXiv, première archive ouverte depuis son lancement, et elle s'est également liée avec RePEC en 2006 et PubMed en 2007. Par conséquent, les identifiants HAL se sont alignés avec les identifiants spécifiques de ces bases pour permettre une interopérabilité et un recoupement optimisé des ressources[248].

L'identifiant HAL identifie chaque ressource mise en ligne et coexiste sur le site avec deux autres identifiants, l'URL permettant le déréférencement ainsi que l'identifiant générique OAI. L'accès se fait systématiquement en ajoutant devant l'identifiant le préfixe « http://hal.archives-ouvertes.fr/ ».

La syntaxe est composée de trois éléments[249] :

- Un préfixe systématique : soit hal, soit tel pour les thèses ou cel pour les cours,
- Un numéro séquentiel de dépôt,
- Un numéro de version indiqué en toutes lettres : version N.

Exemple : hal-123456789 – version 1

En outre, HAL développe son propre identifiant d'auteur, singularisant les chercheurs vis-à-vis de leur authentification lors de leurs connexions sur le site. Cet identifiant, nommé idHAL, fait partie de l'identité numérique du chercheur, et elle est alignée avec les autres occurrences de la personne sur des sites ou dans des bases d'identifiants tels que IdRef, arXivid, ORCID, ResearcherId, VIAF, et ISNI. Une fois l'utilisateur authentifié dans HAL, il peut effectuer lui-même cet alignement en rentrant des métadonnées personnelles (notamment ses identifiants sur d'autres plateformes). Il peut également lier par la suite ses écrits grâce à la « forme auteur » proposée par le site, répertoriant les articles liés à ce nom d'auteur sur HAL[250].

[247] *Archive ouverte HAL*. Hal.archives-ouvertes.fr [en ligne] https://hal.archives-ouvertes.fr/ [consulté le 24/08/2017]

[248] CAPELLI, Laurent. *Identifiants et référentiels dans l'archive ouverte HAL*. MMSH, Ateliers du Numérique LabexMed. CCSD, 2015. [en ligne] Disponible sur https://hal.archives-ouvertes.fr/hal-01119728/file/HAL-Identifants.pdf [consulté le 28/08/2017]

[249] ARCHIMBAULT, Jean-Luc. *Op. Cit*.

[250] CAPELLI, Laurent. *Op. Cit*.

2.3. LES SYSTEMES DE GESTION D'IDENTIFIANTS

Maintenant que nous avons abordé les quelques identifiants principaux qui existent actuellement, nous allons nous pencher sur des questions un peu plus générales qui concernent la gestion même des identifiants. Si dans ce domaine-là aussi un certain nombre de systèmes existe et se concurrence, leur visée n'est pas vraiment la même, et ils proposent chacun des solutions intéressantes. Nous allons voir d'une part ce que l'on pourrait nommer comme des « méta-identifiants », des identifiants servant à identifier et gérer globalement les solutions d'identification entre elles, puis nous nous pencherons plus particulièrement sur les applications et protocoles de redirection qui participent à ces problématiques d'accès et d'identification des ressources. Enfin, nous ferons un petit point sur l'interopérabilité entre identifiants et comment ces systèmes arrivent (ou, justement, n'arrivent pas) à coexister économiquement.

2.3.1. Des identifiants pour gérer des identifiants

INFO

Né en 2003, le concept INFO est issu de la problématique du *http Range 14* qui a pour but d'identifier les entités physiques et les concepts immatériels via des URL. Il est lié au besoin toujours d'actualité de créer des URI sur le web intégrant les systèmes d'identification déjà existants tels que Dewey, LCCN (*Library of Congress Control Numbers*), Bibcode (Système de données de la NASA), PMID (Identifiants de la bibliothèque de Médecine PubMed)… [251] Actuellement, il y a 29 systèmes d'identification enregistrés dans le registre INFO, parmi ceux-ci des identifiants à vocation globale: ARK, DOI, HDL, mais également des identifiants locaux : BNF, ArXiv, DLF, LC, NLA, FEDORA, etc.[252]

Basés sur l'IETF RFC 4452, les identifiants INFO sont maintenu par NISO. L'enregistrement d'une URI INFO est payante, et peut être faite par n'importe quelle organisation qui maintient des espaces de noms, (et pas seulement par l'autorité de maintenance). Les enregistrements sont régulés par un mécanisme de registre. L'idée est de faciliter le référencement des groupes d'information qui ont déjà des identifiants dans les espaces de nom publics via des URI : identifier les identifiants et surtout intégrer ceux qui ne rentreraient pas dans la syntaxe URI habituellement. L'URI INFO ne propose pas de méthode de résolution globale, c'est à l'organisation attributaire de prendre ses propres dispositions.[253]

La syntaxe est simple, il suffit d'ajouter « info : » devant l'identifiant, par exemple info:isbn:979… Ainsi, elle peut être intégrée directement dans une URI et utilisée dans les systèmes de résolution propres à chaque organisation.

URN

Uniform Resource Name est un type d'identifiant qui est souvent confondu avec les URL et les URI, comme nous l'avons vu dans la première partie de ce mémoire. Ce concept a été développé dans l'optique de ne conserver dans

[251] Info-uri.info. *Notice*. 2010. [en ligne] Disponible sur http://info-uri.info/ [consulté le 06/06/2017]

[252] Info-uri.info. *« info » URI Scheme*. [en ligne] http://info-uri.info/registry/OAIHandler?verb=ListRecords&metadataPrefix=oai_dc [consulté le 06/06/2017]

[253] Ietf.org. *The « info » URI Scheme*. [en ligne] Disponible sur http://www.ietf.org/rfc/rfc4452.txt [consulté le 06/06/2017]

l'identifiant que le « nom » de la ressource, en séparant distinctement sa localisation. Cet espace de nom peut tout à fait se combiner avec des identifiants existants, à la manière d'INFO URI. Cela permet notamment de faire des citations pérennes dans le corps d'un texte sans se baser sur des URL instables. [254] Les types d'espaces de nom liés à l'URN réfèrent à des ressources pérennes qui contiennent parfois leurs propres résolveurs.

La syntaxe est la suivante[255] :

- Le préfixe « urn : »,
- Identifiant de l'espace de nom (NID : *Namespace Identifier*) suivi de « : »,
- Chaine spécifique de l'espace de nom (NSS : *Namespace Specific String*).

Exemple : urn :isbn :978…..

Les espaces de noms qui peuvent être utilisés par l'URN sont disponibles sur le site de l'IANA qui en gère « l'attribution »[256]. Parmi ceux-ci nous retrouvons EBU, EIDR, IETF, MPEG, ISO, ISBN, ISSN, UUID, OAI…

XRI

eXtensible Resource Identifier correspond à un identifiant surnuméraire qui se construit à partir d'URI et d'IRI : en effet, il ajoute des éléments de contexte pour les identifiants abstraits[257]. Il ne définit donc pas de périmètre d'action, il vise à englober un maximum de cas d'identification (personnes, organisation, document, concepts, objets, etc.). Ce type d'identifiant est développé par Oasis de manière *Open Source*, même si cela reste controversé vis-à-vis de problématiques de droits[258]. L'objectif premier est de proposer un format universel pour homogénéiser les identifiants abstraits locaux, globaux, actuels et futurs et avoir une interopérabilité au niveau international. Il propose en outre de concaténer plusieurs identifiants et ainsi créer de nouvelles possibilité d'identification grâce à la précision du référent (co-référencement ou *cross-reference*). L'émergence d'XRI correspond au besoin d'obtenir une identification qui remplit toutes les conditions pour gérer tous ces critères en même temps.

L'idée est d'appliquer un niveau d'englobement qui peut être stratifié : un XRI peut en englober un autre, qui englobe lui-même un IRI ou un URI, etc. Ceci inclut des éléments de contexte « réassignables » par des *Global Context Symbol* (abréviés GCS, par exemple le signe égal, le point d'exclamation…), des métadonnées évolutives sont donc ainsi intégrées à même l'identifiant (@ indique que ce qui va suivre est une organisation, = indique que ce qui va suivre est une personne, + un concept, $ les organismes de standard, etc.) Le principe correspond un peu à une

---

[254] Des exemples d'usage d'URN sont décortiqués de façon simple et ludique dans un article du site Ben Meadow Croft.com. *URNs and bibliographic citations in web authoring*. [en ligne] Disponible sur http://www.benmeadowcroft.com/webdev/articles/urns-and-citations/

[255] Ietf.org. *URN Syntax*. 1997. [en ligne] Disponible sur http://www.ietf.org/rfc/rfc2141.txt [consulté le 07/06/2017] Pour faire cette référence j'aurais d'ailleurs pu utiliser l'URN suivante : « urn :ietf :rfc :2141 » car ietf fait partie des espaces de nom d'urn

[256] Iana.org. *Uniform Resource Names (URN) Namespaces*. 2017. [en ligne] Disponible sur https://www.iana.org/assignments/urn-namespaces/urn-namespaces.xhtml [consulté le 07/06/2017]

[257] Oasis. *XRI 2.0 FAQ*. 2005. [en ligne] Disponible sur https://www.oasis-open.org/committees/download.php/15695/xri-2%200-faq-2005-12-01.pdf [consulté le 22/05/2017]

[258] Voir l'archive Wikipédia sur le sujet: http://archive.wikiwix.com/cache/?url=http%3A%2F%2Fwww.fsf.org%2Fnews%2Foasis.html [consulté le 22/05/2017]

« carte joker » pouvant évoluer avec les besoins.[259] Le mélange de segments pérennes et de segments réassignables rend la structure de l'XRI particulièrement souple.

Par exemple, cela permet d'identifier un ouvrage particulier dans une bibliothèque particulière, en employant au sein de l'XRI correspondant une identification de la structure possédant l'ouvrage, concaténée avec son identifiant ISBN (chose impossible à faire avec un simple URI). De même, une personne pourrait être identifiée avec un XRI qui intègrerait sa page d'accueil de site, son adresse email, son identifiant de messagerie instantanée et ce de manière centralisée ou décentralisée en pouvant tout à fait suivre l'évolution de ces paramètres.

La syntaxe se compose de 5 éléments :

- Le préfixe xri:// (qui n'est pas tout le temps obligatoire par ailleurs),
- L'autorité (par exemple, le nom de domaine),
- Le chemin (exprimé avec des slashs comme séparateurs),
- La requête (commence par un ?),
- Le fragment (commence par un #).

En outre, les caractères * et ! servent à indiquer la réassignation possible de certains éléments (*= réassignation possible, !=pérennité). Ainsi, un XRI qui commence par xri:// !! indique que la suite correspond à un identifiant pérenne non amené à évoluer. Pour la résolution, XRI se base sur le protocole http et des fichiers XML simples. Les XRI sont attribuées par programmation.

WebID

Basé sur une initiative du W3C, le WebID a pour objectif d'identifier toute personne physique ou morale sur le Web afin de développer le Web Social. Cela s'inscrit totalement dans la démarche du web de données, dont l'objectif est de permettre à n'importe qui de dire tout sur tous les sujets[260]. Il s'agit de répondre principalement au besoin de regrouper de manière cohérente les informations partagées relatives à chacun selon son activité en ligne : publications personnelles et professionnelles, CV, post de réseaux sociaux, projets, intérêts, lien avec les autres personnes, etc.

Le système WebID permet entre autres de faire des liens entre les différents identifiants locaux assignés par les réseaux sociaux afin de pouvoir établir des ponts entre les données publiées dans les « silos » de type Facebook, Linkedin, Twitter… Les entreprises de ce type elles-mêmes sont fortement invitées à procurer à leurs utilisateurs des WebID qu'ils pourront éventuellement réutiliser dans leurs autres comptes. [261] Tout cela est géré avec le vocabulaire FOAF, spécialisé dans la description de personnes.

L'objectif est également à terme de pouvoir authentifier ainsi des personnes. Plusieurs projets corrélés se greffent à ce concept : l'authentification au moyen de clés publiques/clés privées mimant les systèmes de signature électronique, ou encore

[259] AYERS, Danny, VÖLKEL, Max. *Op. Cit.*

[260] Comme nous l'avons vu précédemment, le principe AAA : *Anyone can say Anything about Anything* est un principe fondateur du web sémantique imaginé par Tim Berners-Lee. Le Web social et le Web de données sont donc complémentaires vis-à-vis de cet objectif.

[261] W3.org, *WebID.* [en ligne] Disponible sur https://www.w3.org/wiki/WebID [consulté le 06/06/2017]

le fonctionnement comme un « *méta-log-in* » (identifiant unique racine pour tous les comptes) pour l'ensemble des activités effectuées sur le web, sécurisé par certificat. Les WebID sont cependant à l'heure actuelle encore très peu utilisés.

OpenID

Toujours dans l'identification des personnes, une autre initiative a vu le jour initiée par la fondation OpenID, organisation à but non lucratif comprenant de grands acteurs du web tels que Yahoo, Google, Myspace.... L'identifiant OpenID est utilisé également par Facebook. Il s'agit ici de développer un identifiant unique gratuit permettant l'authentification (connexion, *log-in*) à un ensemble de sites sociaux, comme WebID.[262] L'idée est que le compte créé est authentifié une seule fois : le fournisseur de l'identifiant va s'assurer de la connexion et transmettre le message aux sites que la personne est bien ce qu'elle prétend être tout au long de sa navigation.

Le format de cet identifiant OpenID dépend fortement du fournisseur, parce qu'ils sont plusieurs à en délivrer. Ainsi, un OpenID attribué par Google sera dans un format différent de celui attribué par Wordpress : Google proposera invariablement un OpenID constitué de l'adresse email Gmail tandis que Wordpress favorisera un nom de domaine lié au site de la personne. De même, certains fournisseurs seront plus intéressants que d'autres et proposeront des services plus diversifiés. C'est notamment le cas de Yahoo ! qui propose des mesures anti-*phishing* mais qui n'accepte pas les profils multiples. A l'inverse, Vidoop propose lui de mémoriser certaines informations et d'autoriser les profils multiples mais n'est pas protégé par un mot de passe.[263]

2.3.2. Applications et protocoles de redirection

PURL

Les *Persistent Uniform Resource Locator* sont des identifiants proposés par OCLC (*Online Computer Library Center*) qui permettent d'assurer la pérennité d'une identification sur le principe de l'ajout d'un intermédiaire. Il s'agit en fait d'effectuer systématiquement la redirection d'un usager sur une URL lorsqu'il appelle l'identifiant PURL : l'URL de base peut évoluer mais tant que le lien est conservé avec la table de correspondance des identifiants, l'accès à cette ressource est pérennisé.[264]

S'il y a tout de même un coût dans la maintenance des tables de correspondances de la part d'OCLC (un service de résolution finalement), le code du système est diffusé en open-source afin qu'il soit plus largement utilisé et implémenté dans les structures.[265] C'est le cas du US Government Printing Office

---

[262] Openid.net. *What is OpenID?* [en ligne] Disponible sur http://openid.net/what-is-openid/ [consulté le 06/06/2017]

[263] Openidexplained.com. *Why should I use OpenID?* [en ligne] Disponible sur http://openidexplained.com/ [consulté le 06/06/2017]

[264] Détail intéressant à remarquer : le serveur ayant changé de nom de domaine, lorsque l'on cherche dans notre navigateur favori l'ancienne adresse du résolveur PURL (à savoir http://purl.oclc.org) on est automatiquement redirigé vers la page actuelle de résolution des PURL : https://archivengines.wordpress.com/2012/08/24/systeme-purl/ (le résolveur PURL est donc lui-même identifié en PURL) [consulté le 07/06/2017]

[265] Sites.google.com. Open, persistent identifiers for managing Web resources. [en ligne] Disponible sur https://sites.google.com/site/persistenturls/ [consulté le 07/06/2017]

qui possède son propre résolveur PURL (http://purl.fdlp.gov). En 2016 le résolveur PURL d'OCLC s'est vu transféré à Internet Archive qui en héberge actuellement le système[266].

La syntaxe se base sur le résolveur que l'on souhaite utiliser, par exemple si on utilise celui d'Internet Archive elle se présente ainsi :

- Le préfixe « http:// »,
- L'adresse du résolveur (ex : archive.org/services/purl),
- /le nom attribué à la ressource par l'autorité nommante.

Permalink (Permalien)

Comparé au système PURL, le Permalien est à peu près basé sur le même principe. Il va « figer » une URL, souvent en la raccourcissant (épuration) et en la simplifiant. Parfois, c'est au contraire en la rallongeant que cela permet d'en garantir l'unicité (c'est au gérant du site de s'en préoccuper) en contenant des informations plus précises, et ce afin d'en faire un identifiant pérenne. Ainsi, il y a moins de risques qu'elle évolue. Si la ressource est modifiée ou déplacée, le Permalien la suivra. En pratique, il est moins pérenne que le PURL qui va, lui, permettre une certaine souplesse dans l'identifiant même grâce à la redirection.

OpenURL

OpenURL est une spécification ANSI/NISO datant de 2004. Elle permet la résolution en prenant en compte l'origine de la demande en plus des éléments de la requête eux-mêmes. Cela fonctionne lors de la requête : l'identifiant qui est assigné à un document en particulier est renvoyé au résolveur de lien avec un paquet de métadonnées nommé *ContextObject* contenant les informations sur l'origine de la requête (contexte pouvant comprendre le type de lien, les droits affiliés au compte utilisateur, l'adresse IP du demandeur, etc.). Il s'agit donc d'ajouter des métadonnées d'origine et de description permettant de faire un lien contextuel : la ressource concernée est contextualisée dans la bibliographie qui la cite, ainsi que sur les éléments qui la décrivent, la façon dont elle est utilisée, et le protocole sélectionné. Ces métadonnées sont écrites lors de la requête et envoyées en XML ou en KEV (*Key Encoded Value*).[267]

En fonction de l'analyse que le résolveur aura fait de ce paquet de métadonnées, il pourra renvoyer une résolution adaptée au demandeur en fonction de ses droits d'accès (redirection vers un hébergeur plutôt qu'un autre, affichage sélectif des métadonnées sur le document, entièreté ou seulement une partie du document, etc.). Cela est intéressant pour la gestion fine des droits d'accès sur le web, notamment pour la diffusion de contenus informationnels.[268]

---

[266] Archive.org. PURL Administration. Voir https://archive.org/services/purl/ [consulté le 07/06/2017]

[267] Figoblog.org. *OpenURL : qu'est-ce que c'est ?* 2004. [en ligne] https://figoblog.org/2004/06/04/207/ [consulté le 07/06/2017]

[268] DALBIN, Sylvie, GIRAUD, Odile. L'OpenURL en quelques mots. *Documentaliste-Science de l'information.* ADBS, 2008. Vol. 45. [en ligne] Disponible sur http://www.cairn.info/revue-documentaliste-sciences-de-l-information-2008-2-page-4.htm#s1n6 [consulté le 07/06/2017]

SRU (Search/Retrieve URL)

Le protocole Search/Retrieve via URL est un système fonctionnant avec http qui permet la structuration des requêtes adressées à un catalogue, ainsi que la formulation des réponses de celui-ci. Il est associé à un service de description, ZeeRex, ainsi qu'à un langage de requête particulier, le CQL (*Contextual Query Langage*) fonctionnant uniquement sur les applications en REST. Ce standard Oasis dont la version 1.1 a vu le jour en 2004, comme OpenURL, est maintenu par la bibliothèque du Congrès et il en est à sa 3ème version actuellement.[269] Il est complété par un protocole « jumeau », SRW (*Search/Retrieve Web*) qui lui est plus adapté aux applications web en SOAP (*Simple Object Access Protocol*). Les deux sont abrégés en SRU/W.

L'idée globale est de pouvoir préciser un format autant pour les requêtes que pour la structure des réponses. En effet, celles-ci diffèrent de manière drastique d'un moteur de recherche à l'autre, et l'idée de l'implémentation du système SRU est de permettre la structuration de ces requêtes/réponses dans le but d'accroitre leur efficacité. Il permet notamment un mode de requête synchrone.[270]

Il existe en réalité de trois types de requêtes : « *Explain* » qui permet de renseigner sur les possibilités d'interrogation du serveur en question ; le « *searchRetrieve* » qui est la requête en elle-même formulée en CQL ; et le « scan » qui correspond à la listes des entrées d'un index.[271]

2.3.3. Interopérabilité entre identifiants « concurrents »

Après ce premier grand tour d'horizon de ce qui se présente à nous en termes de possibilité d'identification, nous ne pouvons que constater que les initiatives sont très nombreuses, très variées, et que chacune essaye de « tirer à soi la couverture » afin de récupérer la plus grande part d'utilisateurs. Bien évidemment, dans ce genre d'environnement économique, le plus gros « poisson » récupère les gains. Chacun souhaiterait que son standard soit adopté, mais ceux-ci présentent des failles qui ne peuvent convenir aux spécificités de tout le monde.

Nous avons vu dans ces différentes descriptions que beaucoup de ces identifiants se recoupent et se rejoignent sur l'identification de ressources, contenus, agents. Le cas de l'ISAN et d'EIDR est édifiant sur ce point-là : l'interopérabilité de deux systèmes est possible et même souhaitée.[272] Uniformiser les identifiants relèverait de l'utopie : un seul identifiant global serait peut-être trop large et non adapté aux différents besoins des communautés spécifiques (par exemple, le cas de l'ISNI et son complément ORCID pour les chercheurs), et une multitude d'identifiants proliférant sans se consulter serait désastreux. L'histoire des identifiants semble pourtant se faire dans le bon sens car si uniformiser n'est pas possible, créer de l'interopérabilité entre les systèmes sans en adopter un pour détruire l'autre est une attitude qui profite à l'ensemble des acteurs du domaine. La spécialisation de certains identifiants tout en conservant la capacité de faire des liens

[269] Loc.gov. *SRU/CQL*. [en ligne] Disponible sur http://www.loc.gov/standards/sru/ [consulté le 07/06/2017]

[270] Ariadne.ac.uk. *An introduction to the Search/Retrieve URL Service (SRU).* 2004. [en ligne] Disponible sur http://www.ariadne.ac.uk/issue40/morgan [consulté le 07/06/2017]

[271] Bnf.fr. *SRU (Search/Retrieval via URL).* 2013. [en ligne] Disponible sur http://www.bnf.fr/fr/professionnels/proto_sru/s.proto_sru_intro.html [consulté le 07/06/2017]

[272] Adami, BFI, CEPI, CineRegio, EIDR, et al. *Declaration on audiovisual work identifiers.* [en ligne] Disponible sur https://ec.europa.eu/licences-for-europe-dialogue/sites/licences-for-europe-dialogue/files/9-AV-identification.pdf [consulté le 06/06/2017]

entre chaque identification d'un même référent est la base du web sémantique. La coréférence n'est de toute façon viable que lorsqu'elle est cartographiée (« mappée »), et ces liens entre les identifiants permettent une traduction directe et fluide tant que le dispositif technique est opérationnel. On pourrait faire le parallèle avec les langues humaines : elles cohabitent et évoquent les mêmes idées mais sont traduisibles tout en gardant la richesse des spécificités de chacune.

Prenons par exemple un cas concret : le projet DiVA de l'Université d'Upsalla en Suède, (et en collaboration avec d'autres établissements du Danemark et de la Norvège), a mis en place les URN : NBN (*National Bibliographic Number*). Ce schéma intègre les identifiants DOI et ARK. De même, le *World Data Center for Climate Change* (WDCC) utilise à la fois des identifiants DOI et URN pour nommer les données scientifiques. Un DOI est assigné par l'archive de données et ensuite est envoyée à la DDB (*Die Deutsch Bibliothek*) en URN pour permettre qu'il soit référencé dans le catalogue de la bibliothèque. Le projet EPICUR, quant à lui, vise à remplacer toutes les URL avec des URN maintenues et résolvables pour tous les objets archivés à la DDB. Ainsi, le serveur peut faire des requêtes URN externes à des collègues européens tels que… l'Université d'Upsalla en Suède.[273]

L'attribution d'URN permettant la résolution en URL et inversement se fait en trois étapes à la *National Széchényi Library* : le propriétaire du document envoie une requête au serveur central, le serveur vérifie si le document existe en HTML et ne possède aucune URN, ensuite le demandeur assigne une URN au document, puis la finalisation de la requête se fait en vérifiant que le document contient bien une URN, qu'il est en format HTML et qu'il ne fait pas doublon avec une autre URN. Tout est *Open-source*, et si le document correspond aux critères, l'URN est acceptée dans la base de données centrale.[274]

Les approches hybrides de ce type montrent que plusieurs systèmes d'identifiants pérennes peuvent être implémentés pour convenir à une organisation : il n'y a pas de solution pouvant convenir à tous les usages.

[273] ERPA Seminar. *Op. Cit.*

[274] *Ibid.*

Conclusion de la partie 2

Dans cette partie, nous avons donc vu que :

- L'identifiant idéal possède bien des qualités qui sont parfois difficiles à concilier, voire même quasiment impossible à toutes respecter,
- Les systèmes d'identifiants pérennes sont les bons élèves de cet environnement et tentent de l'approprier un marché difficile à conquérir,
- Il y a des méthodes d'identification pour les objets réels et pour les objets web, même si effectivement la question ne semble jamais trouver de réelle solution stable,
- Les solutions d'identification fourmillent : toutes ont des choses intéressantes à apporter et toutes sont imparfaites, mêmes celles issues des grands consensus internationaux,
- Vouloir homogénéiser se fait forcément au détriment des spécificités qui font la richesse des initiatives individuelles,
- Les identifiants les plus efficaces et les mieux construits s'imposent d'eux-mêmes, même s'ils proviennent de petites structures initialement,
- L'interopérabilité entre systèmes est la clé de l'énigme, même si elle est difficile à mettre en place.

Les identifiants coexistent donc, fusionnent et meurent tels des organismes vivants dans un vrai écosystème économique. Ils sont au cœur des enjeux financiers, sociétaux et culturels liés à la diffusion de contenus sur le web, puisqu'ils apportent à la fois le moyen de d'accéder aux ressources mises à disposition et la possibilité de les utiliser. Cela entraine des questionnements très vastes pouvant toucher à plusieurs problématiques. Tout d'abord des problématiques sociales : qui a vraiment accès à l'information, qui est favorisé, à qui les ressources sont-elles destinées, quelle est la traçabilité des utilisateurs ? Ensuite des problématiques économiques : quels modèles économiques sont utilisés pour faire du chiffre d'affaire sur la mise en ligne des ressources, à qui profite les dispositifs, comment les organisations se placent-elles sur le marché les unes par rapport aux autres, quelles évolutions sont possibles dans ce marché et comment s'y placer ? Et enfin des problématiques techniques : quelles implémentations sont vraiment efficaces, comment s'adapter au mieux à une structure, de quoi sont faites les réussites et quelle expérience tirer des difficultés rencontrées ?

Dans la troisième partie nous traiterons principalement de cette dernière série de problématiques ayant trait à la technicité, tout en permettant des liens avec les aspects économiques, indivisibles du sujet.

3. USAGES ET BONNES PRATIQUES

3.1. ETUDES DE CAS

3.1.1. BnF, *success story* de l'identifiant ARK

Contexte et enjeux

La BnF, ou Bibliothèque Nationale de France, est une institution à l'héritage historique important. A ses origines, se trouve tout d'abord Charles V, qui aménage une première bibliothèque royale de 917 ouvrages. La mise en place du dépôt légal en 1537 sous François I[er] par lequel tous imprimeurs ou libraires doivent déposer un exemplaire de leur production mise en vente apporte un enrichissement constant à cette bibliothèque, qui évolue et se complète. Louis XIV (et surtout Colbert) en 1666 développe drastiquement son rayonnement, en ajoutant les ouvrages importants des savants de leur temps. Progressivement, elle subit des bouleversements et des améliorations, tant par l'apport de bibliothécaires inspirés que de pertes ou d'enrichissements liées aux évènements de la Révolution française. Elle devient par cette dernière bibliothèque nationale et s'institutionnalise.

La Bibliothèque nationale de France telle que nous la connaissons aujourd'hui est un complexe de bâtiments au cœur de Paris dédiés à la connaissance, la conservation et l'exploitation des collections. Elle émerge sous l'impulsion de François Mitterrand en 1988, qui prend acte des problèmes de stockage et de l'explosion des productions imprimées courant du XXe siècle pendant et après-guerre. Le décret actant sa création est signé en 1994, qui en définit les missions en détail, et sa construction s'achève réellement en 1998.[275]

> « La BnF a pour mission de collecter, conserver, enrichir et communiquer le patrimoine documentaire national […] »[276]

Ces missions primordiales, touchant actuellement à la gestion d'un nombre de documents s'élevant à 15 millions, dont précisément 4 608 377 documents numérisés parmi lesquels 4 008 038 consultables en *open-access* sur le web[277], s'intègrent dans un paradigme numérique qui nécessite la mise en place de systèmes d'information. Actuellement, la BnF possède un budget global s'élevant à 232 millions d'euros, dont 50 millions dédiés au fonctionnement, 223 aux ressources et 43 aux investissements. De grands projets de numérisation, de mise en ligne de ressources, d'archivage électronique et de gestion informatique des catalogues se développent afin de franchir le pas du numérique et proposer de nouveaux types d'accès aux collections. L'intégration de celles-ci à partir des années 2000 à l'espace numérique grandissant (web, systèmes, bases de données) est une priorité pour la BnF afin de mener à bien ses missions et tirer le meilleur parti de ces nouveaux outils.

Dès lors, la définition d'identifiants s'est révélée être une question cruciale qui a intéressé des équipes et groupes d'études spécifiques. Ceux-ci ont bien documenté

---

[275] BnF. *Histoire de la BnF*. Site bnf.fr [en ligne] Disponible sur http://www.bnf.fr/fr/la_bnf/histoire_de_la_bnf.html [consulté le 01/08/2017]

[276] BnF. *Les missions de la BnF*. Site bnf.fr [en ligne] Disponible sur http://www.bnf.fr/fr/la_bnf/missions_bnf.html [consulté le 01/08/2017]

[277] Chiffres correspondant au recensement des collections du 31 décembre 2016 par la BnF, dont l'infographie est disponible sur http://www.bnf.fr/fr/la_bnf/bnf_en_chiffres.html [consulté le 01/08/2017]

leurs travaux et leurs avancées et mis à disposition leurs retours d'expérience. Nous proposons d'étudier ici leurs résultats.

La recherche de l'identifiant idéal

La BnF gère donc des documents très hétérogènes qui ont potentiellement chacun leur propre identifiant en fonction de leur type, qu'ils soient sous la forme d'un ISBN, d'une cote Dewey, d'une URI, d'une DOI, etc. Il y a, de plus, de multiples numérisations qui entrainent la génération d'identifiants spécifiques pour les documents, avec une granularité assez importante (page par page souvent). En outre, les processus de numérisation et de collecte sont très variés, ce qui augmente encore l'hétérogénéité des identifiants gérés à la BnF.

> « Chacun de ces processus dispose nécessairement de ses propres identifiants de production, qui logiquement sont tous différents puisqu'ils répondent à des besoins différents. »[278]

Cette prolifération n'est pas gérable au niveau global, il a donc fallu à la BnF trouver un identifiant pouvant être un repère « pivot » au centre du système.

Dans la recherche d'un identifiant pouvant prendre en compte ses contraintes spécifiques, la BnF a envisagé certains des systèmes que nous avons développés dans notre seconde partie. C'est le cas notamment du système DOI, écarté en raison du prix par identifiant attribué. Le volume très étendu des ressources à identifier ainsi que son expansion continue aurait en effet généré un coût global excessif. PURL et Handle ont également été écartés en raison de la dépendance technique à un service de maintenance. De plus, la BnF souhaitait au maximum éviter l'accumulation des redirections là aussi en raison de la maintenance drastique que nécessiterait une masse de données aussi gigantesque.[279] URN a cependant été un candidat plus sérieux car il était facilement implémentable sur un serveur web local, ne nécessitait pas une infrastructure globale construite spécifiquement pour lui, et il était assez précis sur certains champs (utilisation de qualifiants, politique de pérennité, etc.)[280]. Au terme d'une étude approfondie, la BnF a néanmoins choisi le système d'identification ARK, au vu de l'intérêt qu'il pouvait présenter sur tous les critères cités ci-dessus ainsi que de sa souplesse effective.

Les identifiants ARK à la BnF

L'utilisation d'identifiants ARK est lancée à la BnF en 2006. Cette dernière constitue l'une des structures pilote majeures permettant le retour d'expérience sur leur implémentation à grande échelle, parmi 270 institutions culturelles de par le monde. La perspective est évidemment le très long terme, puisqu'il s'agit d'une institution à objectif patrimonial très fort.

Le système ARK a actuellement 13 ans d'ancienneté, et est issu (et maintenu) par la California Digital Library (CDL). Le président de la BnF de l'époque a signé lors de l'accord avec la CDL un *Memorandum of Understanding* quand la BnF est

---

[278] Ministère de la Culture et de la Communication. *Op. Cit.*

[279] PEYRARD, Sébastien, TRAMONI, Jean-Philippe, KUNZE, John. The ARK Identifier Scheme : Lessons Learnt at the BnF and Questions Yet Unanswered. DC-2014 *Metadata Intersections: Bridging the Archipelago of Cultural Memory*. International Conference on Dublin Core and Metadata Applications, USA, 2014. [en ligne] Disponible sur http://dcevents.dublincore.org/IntConf/dc-2014/paper/view/241

[280] *Ibid.*

devenu officiellement une *Name Mapping Authority*[281], pour officialiser les engagements de pérennité liés à l'utilisation des identifiants ARK. C'est en effet une caractéristique importante du concept d'ARK : faire en sorte que les NMA soient explicitement impliquées dans des objectifs de pérennité de leurs ressources, et que cet engagement soit accessible et lisible par les utilisateurs des ressources identifiées en ARK.

Comme DOI ou URN, ARK a l'avantage d'être indépendant des protocoles DNS et http. Il présente les bons points d'URN (la nomination pérenne et unique d'une entité, que l'on appelle le *core immutable ID**, ou cœur immuable de l'identifiant) avec un déréférencement possible par une conversion en URL assez facile. Celui-ci est d'ailleurs possible à travers plusieurs hébergements : une ressource peut se retrouver avec deux URL qui font référence au même identifiant et à la même ressource, ce qui permet d'avoir des ressources accessibles par différents sites tout en étant globalement uniques.

ARK introduit dans son système la notion de *qualifieur**. Il s'agit de petites chaînes de caractères ajoutées à la suite de l'identifiant ARK qui ne feront pas réellement partie de l'identification pérenne mais qui permettront entre autres : le *versioning* des entités, la gestion de services non liés à des contraintes de pérennité, la granularité des documents plus en finesse, et la définition des formats d'encodage.

ARK propose en outre des services, tels qu'une alternative à la négociation de contenu ou des accès particuliers selon les suffixes apposés à l'identifiant de base. Par exemple, l'ajout d'un point d'interrogation à un identifiant ARK en URI http permet d'avoir accès aux métadonnées du dit-document, tandis que l'ajout de deux points d'interrogation donne accès à la politique d'engagement de l'autorité nommante (en l'occurrence, la BnF) vis-à-vis de la pérennité des ressources qu'elle propose.

De plus, la BnF assure la sécurité de la conservation de ses données en confiant un doublon à la National Library of Medecine aux Etats-Unis, arrangement qui est réciproque.

Méthodes d'implémentation et d'assignation

Initialement, les identifiants ARK ont commencé à être implémentés sur les ressources numérisées de Gallica, la bibliothèque numérique en ligne de la BnF, puis, progressivement ils se sont étendus aux notices de catalogue pour permettre notamment le moissonnage.[282] L'implémentation de SPAR (*Scalable Preservation and Archiving Repository*) pour l'archivage électronique en 2010 a permis de donner une impulsion à leur développement ainsi qu'une étendue supplémentaire, en délivrant systématiquement des identifiants ARK aux ressources lors de leur intégration dans le système si celles-ci n'en possèdent pas encore. SPAR est d'ailleurs actuellement la méthode centrale d'assignation d'identifiants ARK pour les objets digitaux. Il y a néanmoins d'autres procédures d'attribution notamment pour les notices bibliographiques, qui ne sont pas concernées par ce système.

Les méthodes d'attribution peuvent être de trois types[283] :

---

[281] Ou autorité nommante, voir partie 2.2 pour plus de détails.

[282] PEYRARD, Sébastien, TRAMONI, Jean-Philippe, KUNZE, John. *Op. Cit.*

[283] *Ibid.*

- L'assignation automatisée à partir d'un identifiant déjà créé, que l'on transforme en ARK. C'est le cas des identifiants utilisés pour Gallica, le catalogue et la base de données de recherche Aids. Les ressources possédaient déjà des identifiants fiables préexistants, il a donc suffit d'un simple ajout de caractères en préfixe pour l'appartenance au catalogue principal ainsi qu'un caractère de contrôle en suffixe,
- L'assignation automatisée sans aucun identifiant de base. Cela concerne souvent des ressources de taille moyenne qui n'ont pas encore d'identifiant. Lors de leur versement dans SPAR elles reçoivent ainsi leur première identification, sous forme d'ARK. Le nombre attribué n'est pas totalement opaque car il intègre des éléments d'organisation ou des éléments techniques tels que la sous-catégorie, le nom de l'entrepôt où est stockée la ressource, etc. Cela permet d'assurer l'unicité au niveau des lieux de stockage même,
- L'assignation semi-automatique. Elle concerne un plus petit ensemble de données qui seront assignées par les conservateurs en fonction d'éléments spécifiques.

Les anciens identifiants, eux, sont enregistrés dans les métadonnées, permettant une rétroactivité des recherches par les anciens points d'entrée. Cependant, il y a tout de même parfois une réutilisation qui est faite dans la partie même de l'identifiant pivot, dans le cas où sa pertinence est avérée.

« En l'occurrence, la BnF a fait le choix d'utiliser des identifiants opaques alphanumériques, qui intègrent en partie les anciens codes à barres qui servent à gérer la production de numérisation. »[284]

Le projet data.bnf.fr

Le projet *data.bnf.fr* correspond à la volonté de la BnF de s'inscrire dans le mouvement de l'*open-data*, mis en ligne en juillet 2011. Il fait donc partie des projets qui ont pu voir le jour en profitant des avantages apportés par l'implémentation des identifiants ARK à la BnF. Il réunit pour mettre en ligne des données structurées de types et de formats très variés tels que Intermarc, XML-EAD, DC, etc. Celles-ci sont exprimées en RDF et disponibles en RDF-XML, RDF-N3 et RDF-NT. Outre via la négociation de contenu, ces représentations sont accessibles en ajoutant à l'identifiant des suffixes (par exemple /rdf.xml, /rdf.nt), ou encore via des boutons spécifiques en accès direct depuis les pages.[285]

Certaines données sont liées à des bases externes, telles que id.loc.gov pour les langages et nationalités, dewey.info pour les thèmes, DCMI pour les typologies de documents ou encore DBpedia et VIAF. Le modèle de données utilisé est basé sur le standard FRBR, qui conditionne l'organisation des données en « blocs » : les œuvres, les auteurs et les sujets. En outre, il propose une typologie conceptuelle à 4 niveaux : l'œuvre, l'expression, la manifestation et l'élément. Cela permet notamment de lier les données entre elles et d'effectuer les alignements nécessaires à l'exploitation des données liées :

[284] *Ibid.*

[285] BnF. *Semantic Web and data model*. Data.bnf.fr, 2017. [en ligne] Disponible sur http://data.bnf.fr/en/semanticweb#Ancre3 [consulté le 19/05/2017

> « Lorsqu'une manifestation est explicitement liée à un auteur dans la notice, et quand l'identifiant de cette manifestation est exactement le même que le titre de l'œuvre, alors la manifestation est alignée avec l'œuvre. »[286]

Les alignements sont surtout faits grâce à un algorithme qui détecte les titres qui se recoupent. Des outils de structuration de données embarqués apportent également des métadonnées supplémentaires utilisables dans des cadres spécifiques (Schema.org pour les représentations HTML et OpenGraph Protocol pour les réseaux sociaux). De plus, le site réutilise au maximum les ontologies disponibles afin de maximiser l'interopérabilité des systèmes (et *in fine* la réutilisation des données), telles que SKOS, FOAF, Dublin Core et RDVocab. Il possède néanmoins une ontologie propre comportant 12 propriétés qui lui sont spécifiques, « bnf-onto », qui comporte notamment une propriété définissant les « textes alternatifs » pour les images ou les URL d'une exposition virtuelle.[287]

Déréférencement et accès à la ressource

Dans l'objectif d'un management SEO (*Search Engine Optimization*[288]) des identifiants, la BnF souhaitait que le déréférencement de ses identifiants ARK opaques sur les différents sites se fasse systématiquement via une redirection sur des URI non opaques temporaires permettant la visibilité et la maniabilité sur le web. Ainsi, Sébastien Peyrard, Jean-Philippe Tramoni et John Kunze, travaillant sur le projet, donnent l'exemple suivant : l'identifiant « ark:/12148/cb118905823 », une fois déréférencé devient dans la barre d'adresse du navigateur « /11890582/charles_baudelaire »[289]. Cette URL propose en outre de la négociation de contenu permettant l'accès à différentes représentations du contenu en HTML, RDF, et autre. Il y a également une scission entre le visionneur de document et le résolveur : comme celui-ci ne prend plus en charge la résolution, il est possible de rediriger l'usager vers les applications spécifiques (sous-domaines) de manière plus flexible, et ce à l'aide de trois modules : le premier vérifie que la requête correspond au domaine, et si ce n'est pas le cas il l'aiguille vers le module de redirection ; le deuxième analyse la requête et la reformule avant de la transférer à l'API concernée ; le troisième est le « module de redirection ARK » à proprement parler qui analyse l'identifiant ARK pour le renvoyer à l'API d'un domaine spécifique. Les paramètres de redirection sont définis dans un fichier XML. Pour résumer, la requête est filtrée en premier lieu par les deux premiers modules. Si l'identifiant ARK est reconnu par le domaine, il est tout de suite envoyé aux applications. Si l'identifiant ARK n'est pas reconnu, il est envoyé au module de redirection centralisé qui va l'analyser.[290]

La numérisation fait de plus une grande utilisation des qualifiants (ou qualifieurs), qui permettent de donner des détails techniques ou pratiques facilitant la manipulation humaine : des caractères explicitant si l'objet est un aperçu, une version haute résolution ou basse résolution, une page ou un groupe de pages, une

---

[286] *Ibid.*

[287] *Ibid.*

[288] Pour plus de détails sur cette notion, se reporter à la première partie de ce mémoire.

[289] PEYRARD, Sébastien, TRAMONI, Jean-Philippe, KUNZE, John. *Op. Cit.*

[290] *Ibid.*

version lue ou une version océrisée[291], etc. En ce qui concerne la négociation de contenu, la BnF est en mesure, en fonction des serveurs fournissant l'accès (les NMAH[292]), de sectoriser les représentations qu'elle délivre. En effet, l'on pourrait trouver sur www.catalogue.bnf.fr un format traditionnel de la ressource et sur www.data.bnf.fr sa version HTML ou RDF.[293]

Mais qu'en est-il du déréférencement des objets réels ? La question s'est longuement posée, à l'instar des débats animés au sein de la communauté professionnelle du web sur le sujet. Le problème vient principalement des qualifiants ajoutés aux formes URL. Si à la base l'idée était d'utiliser la redirection 303 comme nous l'avions évoquée précédemment, en pratique c'est plutôt la méthode Hash qui a été employée via une extension locale, car le dièse n'est pas un caractère réservé dans ARK. Le problème est que le définir en tant que qualifiant pouvait également créer des soucis de compatibilité avec les identifiants ARK déjà assignés.[294]

Problèmes rencontrés

Malgré un ensemble de conditions optimales semblant être réunies en début de projet, certains problèmes fonctionnels ont été rencontrés, notamment des soucis techniques vis-à-vis de la résolvabilité de certains ARK. La majorité des problèmes venaient cependant des API, certaines applications conservaient des redirections d'identifiants ARK obsolètes, enregistraient des ARK résolvables uniquement en tant que métadonnées, etc.[295] De plus, quand la ressource n'est pas disponible en URL, la réponse du navigateur est toujours 404 ou 403, alors qu'il faudrait plutôt qu'elle renvoie des informations de type « ressource non trouvée, ressource supprimée, accès refusé, etc. »[296]. Ironiquement, le standard ARK ne donne pas la possibilité aux machines d'interpréter les données rapportées par l'inflexion « ?? » censée afficher l'engagement de pérennité du NMA. Cela pourrait constituer une bonne piste de développement.

Deux autres problèmes sont survenus [297]:

- Celui de la « requête via citation » : Lorsque l'on effectue une recherche spécifique sur un mot précis dans un document, les caractères « .r=mot » s'ajoutent à l'URL, ce qui créé donc par conséquent une nouvelle URL. Celle-ci n'identifie plus vraiment le document mais le document modifié par le surlignage du mot recherché. Les usagers sont donc potentiellement susceptibles de faire des citations à partir d'URL erronées.
- Celui du « technique vs non-technique » : En principe, nous l'avons vu, les qualifiants exprimant des éléments en rapport avec l'implémentation technique ne devraient théoriquement pas se retrouver dans l'identifiant. Cependant

[291] L'océrisation, ou OCR (*Optical Character Recognition*) correspond à des solutions logicielles reconnaissant automatiquement lors de la numérisation d'un texte papier les caractères présents, permettant ainsi un traitement informatique et une récupération des données (traitement de texte, copier-coller, etc.).

[292] NMAH : *Name Mapping Authority Host*, que nous avions évoqué dans la partie 2 de ce mémoire, et qui fait référence à l'hébergeur du contenu mis à disposition par l'autorité nommante.

[293] PEYRARD, Sébastien, TRAMONI, Jean-Philippe, KUNZE, John. *Op. Cit.*

[294] *Ibid.*

[295] *Ibid.*

[296] *Ibid.*

[297] *Ibid.*

contrairement à l'identifiant ARK simple, les qualifiants ne sont pas vraiment conçus pour la préservation sur le long terme, mais ils sont très pratiques concernant la maniabilité pour les usages humains.

Résultats

Les politiques d'identification, très strictes au départ, se sont assouplies au fur et à mesure.[298] La BnF se demande aujourd'hui si elle doit étendre les caractéristiques d'ARK afin d'augmenter l'interopérabilité croisée des résolveurs (en étendant l'utilisation des identifiants ARK à d'autres contextes, par exemple) ou bien rester sur l'implémentation simple et fluide actuelle.

Sébastien Peyrard, Jean-Philippe Tramoni et John Kunze identifient cinq problématiques qui restent encore en suspens à l'issue de ce projet : qui doit prendre les décisions lorsqu'il faut ajouter de nouveaux types d'objets à identifier ? Qu'est-ce que les identifiants identifient réellement (ressource, représentation, nom…)? Est-ce que à terme les identifiants pourront être réassignés ? A quelle quantité de changements au niveau des contenus doit-on s'attendre (et peut-on supprimer des contenus)? Quels services et sous-domaines est-il souhaitable de révéler aux utilisateurs (révéler la mécanique interne) afin de leur permettre d'évoluer dans la granularité des ressources mises à disposition ?[299]

Nous constatons donc que les problématiques rencontrées par la BnF lors de la mise en place de ce projet montrent des préoccupations touchant plus à la gouvernance qu'à la technicité, malgré qu'elles soient elles-mêmes très « pratiques ». La BnF semble avoir très bien géré l'implantation de ses identifiants sur la phase de conception du projet et se place actuellement comme acteur majeur dans l'exploration des fameux points de débat qui animent la communauté. Nous pouvons donc reformuler les problématiques posées par Peyrard, Tramoni et Kunze de la manière suivante :

- La crise identitaire : qu'identifie-je ?
- Le tristement connu http Range 14 : comment représenter les objets réels ?
- La responsabilité et la source de pérennité sur le long terme : sur qui repose les enjeux de pérennité, sur qui porter la responsabilité d'un éventuel échec et à qui se référer ?
- Les problématiques de droit et de confidentialité : jusqu'où doit-on être transparent ? Jusqu'où pousser l'*open-access* ?
- L'optimisation du confort utilisateur (et ainsi, la popularité du service via les algorithmes de pagerank des moteurs de recherche) : comment rendre notre service efficace, intuitif et manipulable ?
- Et enfin, la gestion sur le long terme de l'évolution des ressources (modification, suppression, réassignation, etc.) : serons-nous amenés à réviser les principes d'identification jusque-là inviolables pour nous adapter aux changements, à savoir la réattribution, la suppression et la modification d'identifiants, principes qui touchent aux caractéristiques essentiellement intrinsèques de l'identifiant, l'unicité, l'accessibilité, la citabilité ?

[298] *Ibid.*

[299] *Ibid.*

Le retour d'expérience de la BnF sur ces questions sera d'autant plus intéressant à l'avenir quand le temps aura mis à l'épreuve sur le long terme le système, qui est pour l'instant relativement jeune.

3.1.2. BBC, réutilisation contrôlée

Contexte et enjeux

La BBC (*British Broadcasting Corporation*) est un média anglais qui est apparu en 1922. Elle fut supervisée jusqu'au 1er janvier 2017 par la BBC Trust, un organe de décision prélevant notamment la redevance audiovisuelle des foyers anglais, qui nommait son directeur et dont les membres étaient sélectionnés par la Reine en personne (au 1er janvier la BBC Trust a été supplantée par le Board of the BBC). La BBC a détenu le monopole de la télévision anglaise jusqu'en 1955, juste avant l'apparition d'ITV et des radios locales. Autorité administrative indépendante, son rôle est de diffuser des contenus principalement télévisuels et radiophoniques. Ses productions ont une place privilégiée au sein des chaînes de TV anglaises (de 1000 à 1500 programmes diffusés chaque jour[300]) sur 17 chaînes principales, et elle jouit d'une réputation « d'excellence culturelle ».

En termes de revenu brut et d'audience elle est la plus importante société de diffusion au monde [301]. Segmentée en 4 services principaux, télévision, radio, site web (BBC Online) et vidéo à la demande (BBC Red Button), elle compte 13 départements thématiques. Elle possède en outre 13 filiales comprenant des orchestres, des studios de production, radios, sites web et chaînes de news. Son effectif avoisine les 25 000 employés et elle plafonne à un chiffre d'affaires d'environ 4 milliards 7 de £. Concernant la partie radio, la BBC compte 10 radios nationales et 40 radios locales. Enfin, pour la partie web, elle possède un grand site internet tentaculaire, BBC Online, qui propose des informations et des contenus rediffusés via des microsites dédiés à des domaines spécifiques.[302]

La BBC est donc un gigantesque producteur de contenus au quotidien, contenus audiovisuels qui, d'une part, peuvent être très lourds numériquement (uniquement pour l'archive radiophonique du département BBC World Service, la totalité s'élève à 3 ans non-stop de bande et 15TB de données[303]) et qui, d'autre part, nécessitent une gestion, c'est à dire qu'il faut pouvoir retrouver, relier, rediffuser, conserver et organiser. Le caractère impérieux de cette gestion, augmentant avec l'expansion de la production et de la firme au fil des années présente deux aspects incontournables : une gestion rigoureuse en interne vitale au fonctionnement de l'institution, ainsi qu'une diffusion en ligne du catalogue fournissant des informations sur les ressources aux utilisateurs (voire une plateforme de visionnage en différé (*Replay*) des émissions et programmes). Ces deux axes peuvent tout à fait

[300] RAIMOND, Yves. « Les programmes de la BBC tirent avantage du web de données ». Dans Approches documentaires : priorité aux contenus. *Documentaliste-Sciences de l'Information*. ADBS, 2011. N° 48. p57

[301] *BBC*. Wikipédia [en ligne] Disponible sur https://fr.wikipedia.org/wiki/Portail:BBC [consulté le 03/08/2017]

[302] KOBILAROV, Georgi, SCOTT, Tom, RAIMOND, Yves, OLIVER, Silver, SIZEMORE, Chris, SMETHURST, Michael, BIZER, Christian, LEE, Robert. *Media Meets Semantic Web – How the BBC Uses DBpedia and Linked Data to Make Connections*. British Broadcasting Corporation, Londres. Freie Universität, Berlin. Rattle Research, Sheffield. 2009. [en ligne] Disponible sur https://pdfs.semanticscholar.org/5728/393384a3a60e55a72fa3a01d4fc1b258aac2.pdf [consulté le 01/06/2017]

[303] RAIMOND, Yves, FERNE, Tristan, SMETHURST, Michael, ADAMS, Gareth. *The BBC World Service Archive Prototype*. BBC R&D, 2014. [en ligne] Disponible sur www.sciencedirect.com/science/article/pii/S1570826814000535 [consulté le 01/06/2017]

se réaliser au moyen d'outils communs : la gestion rigoureuse des programmes et des contenus, identifiés de manière unique et pérenne, profite à leur mise en ligne et à leur manipulation sur le web. Cela permet ainsi leur citabilité par les utilisateurs, leur exploitation en dehors des canaux de diffusion classique et l'enrichissement de la discussion à leur propos.

L'objectif du site web de la BBC (www.bbc.co.uk) lors de son ouverture était principalement à permettre la diffusion d'informations concernant les programmes voire éventuellement pour certains la rediffusion en *Replay*. Cependant il ne permettait pas encore l'enrichissement des contenus et leur prolongement via le média « web » (critiques, articles) ni les liens entre eux. L'arrivée des projets de *Linked Data* au sein de la BBC modifie les objectifs de ce site et manifeste sa volonté de participer activement à l'enrichissement du web de données.

BBC Online, partie Programmes

L'idée globale lancée avec le service BBC Programmes en 2007 était de créer du lien en HTML entre les différentes parties isolées du site et de croiser les différents domaines. En effet, à l'origine les domaines dans BBC Online étaient bien segmentés : jardinage, maison, news… On ne pouvait pas réellement rechercher de manière globale tout ce qu'avait à offrir le site sur un sujet en particulier. Les liens vers les différents programmes n'étaient pas fiables et la donnée ne pouvait pas être réutilisée dans d'autres contextes.[304] Il fallait donc constituer une base de données pivot hébergée sur le site principal de la BBC, afin de permettre cet enrichissement et cette réutilisation dans différents services proposés par la BBC. L'attribution d'identifiant uniques, pérennes et facilement manipulables se situait au cœur des enjeux dans cette nouvelle section.

L'entité centrale du modèle est le « programme », produit éditorial variable qui peut être soit un épisode, soit une série, soit une marque (de l'anglais *brand*, par exemple Doctor Who est une marque en *crossmédia* qui relie des contenus très divers). En termes de granularité, celle-ci s'étend sur 4 niveaux : les programmes peuvent avoir plusieurs versions, qui elles-mêmes peuvent avoir plusieurs diffusions. Les différentes diffusions de ce programme peuvent avoir des propriétés associées, la diffusion au cours d'un certain service (tranche horaire) et avec certaines disponibilités (disponible en rediffusion ou pas)[305]. A cette granularité de base s'ajoute celle, plus fine, du modèle de segmentation correspondant au chapitrage ou à la recherche plein-texte pour un livre, de sorte que le contenu ne soit pas opaque. Pour un contenu audiovisuel, cela peut concerner[306] :

- La liste des morceaux joués dans un spectacle musical,
- Les passages notables dans un programme,
- Les sous-sections d'un programme,
- Les découpages liés à l'utilisation des ontologies (*Even Set Time*, que nous évoquerons plus longuement par la suite),

[304] RAIMOND, Yves, SCOTT, Tom, OLIVER, Silver, SINCLAIR, Patrick, SMETHURST, Michael. Dans WOOD, David. *Linking Enterprise Data*. Springer, 2010. Chap: Use of Semantic Web technologies on the BBC Web Sites, p263-283

[305] KOBILAROV, Georgi, SCOTT, Tom, RAIMOND, Yves, OLIVER, Silver, SIZEMORE, Chris, SMETHURST, Michael, BIZER, Christian, LEE, Robert. *Op. Cit*.

[306] RAIMOND, Yves, SCOTT, Tom, OLIVER, Silver, SINCLAIR, Patrick, SMETHURST, Michael. *Op. Cit*.

- Les segments de la *timeline,*

Ces segments sont eux-mêmes classifiés afin de permettre l'ajout de métadonnées plus précises sur la description de lieux, de sujets ou de personnes au sein même des ressources.

En termes de recherche, BBC Programmes permet de créer de simples agrégations grâce à des prédicats de catégorie en SKOS. Ceux-ci distinguent deux types de catégories possibles : les genres, et les formats. En outre, le prédicat « sujet » est utilisé afin de lier horizontalement des thèmes plus généraux entre eux (par exemple les toucans, les planètes, etc.)[307]

Développement des identifiants de programmes

La construction des identifiants de programmes devait se faire originellement dans l'optique d'une meilleure lisibilité pour les utilisateurs. Le choix a cependant été fait de ne pas utiliser les titres des programmes ni les dates de diffusion car ceux-ci sont changeants, et surtout peuvent créer de l'ambiguïté et/ou entrer en collision. De plus, cela exclurait potentiellement les contenus non encore programmés ou non diffusés.

Les équipes travaillant sur le sujet ont finalement opté pour des identifiants opaques, permettant d'augmenter l'unicité et de faciliter l'ajout de suffixes pour obtenir de la granularité.[308] Ils sont spécifiques à la BBC, et sont dotés de suffixes techniques permettant l'ajout manuel donnant accès aux différentes représentations. Ils sont en outre construits sur une logique similaire pour chacun des services de BBC Online :

- L'URL de base du site précisant le nom de domaine, ex : http://www.bbc.co.uk/,
- L'espace de noms correspondant à la catégorisation dans le modèle de données, ex : /programmes/,
- L'identifiant opaque, ex : b00cccvg,
- Le suffixe indiquant à quel type de ressource il s'agit.

Ce suffixe peut être d'ordre variable[309] :

- L'absence de suffixe indique qu'il s'agit d'un document à propos du programme,
- Le suffixe « #programme » identifie le programme en lui-même (notons ici l'utilisation des mécaniques de Hash URI),
- Le suffixe de format (.html, .rdf, .mp) indique une représentation disponible dans le format indiqué.

Cette règle s'applique aux versions, aux segments, aux diffusions et aux disponibilités : un tronc commun, complété par l'indication de programme et la granularité. En outre, la négociation de contenu est elle aussi implémentée, c'est-à-dire qu'une personne qui met en favori une ressource sur son ordinateur en HTML peut la retrouver sur son mobile en format XHTML.[310]

---

[307] *Ibid.*

[308] *Ibid.*

[309] *Ibid.*

[310] *Ibid.*

Si le système est propre à la BBC et développé pour elle, des identifiants externes sont néanmoins réutilisés. C'est le cas des identifiants de BDpedia employés depuis 2009 afin de « tagguer » les contenus.[311] Cela permet de lier les concepts grâce à des équivalences en OWL entre l'identifiant BBC et l'identifiant DBpedia (Owl :SameAs), et d'avoir un matériel d'échange réciproque, la base de la BBC enrichissant les contenus de DBpedia et inversement.[312]

Le service BBC Music

BBC Music est un autre service développé dans cette même optique de lisibilité, hébergé sur BBC Online. BBC Programmes et BBC Music sont évidemment très liés. La raison d'être de BBC Music est de pouvoir générer des recommandations spécifiques en termes de musique à un utilisateur en fonction de ses préférences et de ses habitudes de consultation. En général, les algorithmes générant les recommandations sur les sites sont assez obscurs et ne dévoilent pas quel rapport ils ont trouvé entre la ressource de départ et la recommandation. Ici le principe est de faire en sorte que ceux-ci soient plus transparents. Les liens sur ces données liées se font selon différents degrés de relation, par exemple retrouver des programmes liés à la ville de naissance d'un artiste en particulier, déduire des connexions inédites entre les artistes, etc.[313] La BBC a créé dans ce but deux prototypes d'outils de recommandation : LODations qui permet d'avoir des recommandations relevant d'un intérêt éditorial, par exemple deux groupes de musique formés la même année dans la même ville ; et Musichore, un système de création automatique d'un flux radio avec un DJ « machine » expliquant lui-même via transcription vocale la connexion entre les morceaux qu'il joue.

BBC Music propose donc un guide pour le contenu musical de toute la BBC, permettant de faire le lien entre les artistes, leur musique et les programmes dans lesquels ils apparaissent. Les entités distinguées sont les artistes, les éditions et les labels. Contrairement à BBC Programmes, BBC Music puise ses informations dans plusieurs sites extérieurs qui enrichissent son contenu (et réciproquement) : Musicbrainz pour les informations principales, Wikipédia en ce qui concerne les biographies, et le site de la BBC pour ce qui est des critiques d'albums et des images. En outre, les concepteurs ont beaucoup réutilisé les ontologies disponibles sur la musique, *reviews ontology*, *music ontology* (BBC) et SKOS pour les genres musicaux. [314]

L'identification sur BBC Music est également opaque. Cependant, les identifiants globalement uniques (GUID) de Musicbrainz sont réutilisés pour les artistes, puisque c'est là que sont puisées la majorité de leurs entités. Le système de granularité fonctionne ici comme pour les identifiants de programmes, avec des Hash URI pour les personnes réelles et des suffixes de représentation. En ce qui concerne les critiques d'album, ils ont créé leurs propres URL opaques comportant une « clé » pour chaque critique (par exemple /music/reviews/:url_clé#review). Quant à l'identification de critiques en tant que personnes (journalistes, utilisateur,

---

[311] RAIMOND, Yves, FERNE, Tristan, SMETHURST, Michael, ADAMS, Gareth. *Op. Cit.*

[312] RAIMOND, Yves, SCOTT, Tom, OLIVER, Silver, SINCLAIR, Patrick, SMETHURST, Michael. *Op. Cit.*

[313] *Ibid.*

[314] SINCLAIR, Patrick, HUMFREY, Nicholas, RAIMOND, Yves, SMETHURST, Michael, SCOTT, Tom. *The Web as a Content Management System*. BBC Audio and Music Interactive, 2009. [en ligne] Disponible sur http://www2009.eprints.org/236/1/www2009developers_submission_51.pdf [consulté le 20/06/2017]

etc.) et de contenus mis en avant périodiquement, le principe est le même qu'avec les espaces de nom dédiés /reviewers/ et /promotions/.[315]

Le prototype BBC World Service Archive

Le BBC World Service est un des diffuseurs de programmes radiophoniques de la BBC, qui concerne 28 langues de par le monde. Comme Yves Raimond, Tristan Ferne, Michael Smethrust et Gareth Adams le précisent dans leur article sur le sujet[316], les archives des contenus radiophoniques du BBC World Service ont fait l'objet d'une numérisation massive de 2005 à 2008. Cependant, les données produites par ce service sont très peu exploitées à cause du manque de métadonnées de qualité appliquées aux différentes ressources, les métadonnées ayant été attribuées de manière légère contrairement à d'autres services qui ont bénéficié d'une indexation manuelle[317]. En effet, jusqu'à présent les pratiques qui avaient cours pour la publication en ligne d'archive du BBC World Service étaient systématiquement la segmentation de quelques clips ou programmes qui étaient ensuite correctement traités et décrits avant d'être publiés. L'idée d'un service alternatif de publication automatique des contenus archivés de la BBC émerge du département de recherche et développement de la BBC, qui y voit l'opportunité de développer un outil qui profitera par la suite à l'ensemble du site. L'idée est d'aller plus loin sur l'indexation des contenus, permettre l'intégration des ressources en tant que données liées, mais également par le *crowdsourcing** favoriser une évolution et une amélioration constante des tags et des liens entre elles. Les archives du BBC World Service se prêtent tout à fait à l'exercice, ce qui aboutit à la création d'un prototype.

En premier lieu, le service cherche à créer des liens avec des sources connues, internes ou externes.[318] Il peut par exemple rechercher si des métadonnées sont déjà présentes sur la ressource. Si c'est le cas, il tente ensuite de trouver des associations avec des thèmes identifiés par les IRI des données liées de sources externes, telles que Wikipédia. A ces associations est additionnée une note de « fiabilité ». Les relations sont ensuite traduites en triplets RDF puis liées à DBpedia, tout en constituant des « tags » de métadonnées. Ces liens peuvent également se faire à partir de la piste audio : le système détecte les moments de discours dans l'enregistrement, qui sont ensuite transcrits grâce au logiciel open-source CMU Sphinx. Puis un système de NER « maison » créé avec l'aide du laboratoire de recherche Rattle Research nommé « Muddy Boots »[319] cherche des associations à faire avec les listes d'identifiants de DBpedia. Ce système est intéressant car contrairement à d'autres systèmes de NER, l'association des concepts se fait systématiquement avec les URI de DBpedia[320]. La désambiguïsation se fait en général par le rapprochement de deux tags proches dans la transcription. Les résultats fournis par la prototype sont ensuite

[315] RAIMOND, Yves, SCOTT, Tom, OLIVER, Silver, SINCLAIR, Patrick, SMETHURST, Michael. *Op. Cit.*

[316] RAIMOND, Yves, FERNE, Tristan, SMETHURST, Michael, ADAMS, Gareth. *Op. Cit.*

[317] Précisons que, comme le mentionnent les auteurs, le temps d'indexation manuelle par un archiviste peut être de 9h pour un programme de 30min, et dans le cas des programmes radiophoniques de BBC World Service, 19 000 programmes n'ont même pas de titre et 17 000 n'ont pas de synopsis.

[318] RAIMOND, Yves, FERNE, Tristan, SMETHURST, Michael, ADAMS, Gareth. *Op. Cit.*

[319] KOBILAROV, Georgi, SCOTT, Tom, RAIMOND, Yves, OLIVER, Silver, SIZEMORE, Chris, SMETHURST, Michael, BIZER, Christian, LEE, Robert. *Op. Cit.*

[320] Citons par exemple les systèmes de NER OpenCalais, Zemanta et Twine qui, jusqu'à 2009 en tout cas, utilisaient leurs propres identifiants, ce qui obligeait les utilisateurs à aligner par la suite les concepts.

soumis à l'épreuve de retours par *crowdsourcing* (les utilisateurs peuvent voter pour la pertinence de tel ou tel tag associé à la ressource), et ils sont également comparés à un échantillon de métadonnées produites depuis le départ par des archivistes professionnels (une sorte de *Gold Standard Corpus* en quelque sorte). [321]

Ici nous voyons bien que la pratique d'identification permet la liaison de concepts entre eux et l'indexation automatique de ressources, ce qui représente une réussite malgré l'évidente marge de progression constatée. La BBC utilise les ressources d'acteurs externes pour enrichir et développer les métadonnées de ses propres contenus, qui seraient trop complexes à indexer seuls.

Le service Wildlife Finder

Wildlife Finfer quant à lui constitue une base de données liant les sujets lié à la biologie, l'environnement et le règne animal : le modèle est construit autour de « l'espèce » et sont attribués des identifiants uniques et pérennes à toutes les espèces, les habitats et les adaptations. Les ontologies développées pour ce service sont au maximum interopérables avec les ontologies plus spécialisées des domaines scientifiques (taxonomie, écologie, science environnementale et bio-informatique)[322]. Ici l'idée principale est d'identifier les objets réels, car c'est la raison d'être de ce service, et non pas des contenus en particulier.

Les sources des données présentes sur le BBC Wildlife Finder sont variées : Wikipédia, WWF's Wildlife Finder, la liste rouge des espèces menacées de l'IUCN, le programme EDGE de la Société Zoologique de Londres, etc. Ici le principe consiste en l'intégration de données externes sur le site de la BBC avec l'ajout de métadonnées reliant avec des triplets RDF les concepts entre eux (articles de news, programmes, musique, etc.)[323].

En termes d'identification, Wildlife Finder réutilise une partie des identifiants de Wikipédia (dernière partie de l'identifiant qui correspond réellement à l'identification de la ressource en elle-même), ce qui évite entre autre de maintenir un vocabulaire contrôlé et permet le partage global d'une définition de la ressource. En outre, DBpedia est également utilisé pour son vocabulaire contrôlé. L'identifiant d'un élément biologique apparait donc sous la syntaxe suivante : /nature/rank/ :identifiant-wikipedia# :rank/[324]. En ce qui concerne l'accès aux différentes représentations, comme pour BBC Programmes, cela peut se faire via les suffixes ou via la négociation de contenu.

Problèmes rencontrés

Lors du retour d'expérience effectué en 2009 deux ans après le lancement des services Programmes et Music, les équipes en charge du projet constatent que ceux-ci ne répondent pas réellement à la problématique des liens trans-sectoriels, ni à celle de la désambiguïsation des vocabulaires contrôlés multiples[325]. En effet, un artiste qui serait à la fois un chanteur, un acteur, et une personnalité en soi ne peut

[321] RAIMOND, Yves, FERNE, Tristan, SMETHURST, Michael, ADAMS, Gareth. *Op. Cit.*

[322] RAIMOND, Yves, SCOTT, Tom, OLIVER, Silver, SINCLAIR, Patrick, SMETHURST, Michael. *Op. Cit.*

[323] *Ibid.*

[324] *Ibid.*

[325] KOBILAROV, Georgi, SCOTT, Tom, RAIMOND, Yves, OLIVER, Silver, SIZEMORE, Chris, SMETHURST, Michael, BIZER, Christian, LEE, Robert. *Op. Cit.*

pas être recoupé comme étant une seule et même entité. De plus, ces services ne répondent pas non plus au besoin de lier les contenus avec ceux des autres domaines traités par la BBC (bien que correctement liés entre eux) : la cuisine/l'alimentation, les livres, les news, etc. Le besoin est donc clair : s'accorder sur des vocabulaires contrôlés communs identifiés globalement afin de faciliter les échanges. Après des recherches ayant impliqué plusieurs autres acteurs, notamment la Freie Universität de Berlin, un consensus s'établit autour de l'utilisation des données de DBpedia pour réaliser ces liens.[326]

Une autre problématique rencontrée au cours du projet semble avoir été celle de la gestion du contenu non structuré. En effet, si produire des données liées sur des contenus correctement modélisés et possédant des identifiants uniques et pérennes requerrait déjà un déploiement de méthodes d'automatisation plus ou moins complexes, le faire sur du contenu non structuré a été semble-t-il autrement difficile. Les recherches effectuées ont permis de créer des pages d'agrégations de contenus non structurés et de contenus structurés (les deux permettant de faire des recoupements) afin de les catégoriser de manière automatique au moyen là encore des identifiants et concepts de DBpedia. [327]Ces pages ont ainsi été employées comme des outils de médiation pour l'indexation de ces contenus. A mesure de l'utilisation faite et des liens créés, l'équipe éditoriale retire ou ajoute les liens de redirection.

Avec Wildlife Finder, la problématique a surtout été liée au choix des classifications, car nous l'avons vu le domaine de la taxonomie est très subjectif[328]. La compréhension actuelle que l'on a de la biologie et du règne animal est en constante évolution, les points de vue peuvent varier, la structure est nécessairement hiérarchique et donc n'est pas absolue. La seule définition sur laquelle les équipes pouvaient se baser est celle de « l'espèce », qui est le pivot global de ce service. Ils ont en outre fait le choix de définir chaque espèce dans une classe « espèce », et non chaque espèce constituant une classe distincte.[329] Les choix qui ont été faits en termes d'organisation ont systématiquement été personnalisés en fonction du service, ce qui a demandé un travail d'analyse conséquent en phase de préparation du projet, et beaucoup de tests utilisateurs.

Résultats

Nous constatons que la BBC a développé une stratégie de gestion d'identifiant adaptative : elle a fait le choix de créer pour ses contenus propres des identifiants « maison », et, dans les cas de réutilisation de données externes, de conserver les identifiants d'origine en les intégrant dans le *scheme* de ses identifiants. D'une part, cela lui permet d'interopérer d'un côté comme de l'autre (vers l'intérieur ou vers l'extérieur) en gardant une globalité et une homogénéité interne entre ses identifiants, d'autre part cela démontre une forte volonté de s'imposer dans l'identification de sa production et d'en définir elle-même les critères. Ce *melting-pot* d'identifiants et de sources est utilisé à son avantage tout en proposant systématiquement des retours gagnant-gagnant avec ses partenaires.

---

[326] *Ibid.*

[327] KOBILAROV, Georgi, SCOTT, Tom, RAIMOND, Yves, OLIVER, Silver, SIZEMORE, Chris, SMETHURST, Michael, BIZER, Christian, LEE, Robert. *Op. Cit.*

[328] Voir la partie 1.2 pour plus de détails.

[329] RAIMOND, Yves, SCOTT, Tom, OLIVER, Silver, SINCLAIR, Patrick, SMETHURST, Michael. *Op. Cit.*

Entre-autres, la BBC a su de ce fait exploiter les identifiants des autres structures au-delà de la simple réutilisation, en optimisant l'indexation de ses contenus avec des méthodes développées en interne qui utilisent les avantages apportées par les données liées. Elle participe également à un web plus riche en permettant aux organismes qui s'impliquent dans des échanges de ce type de réutiliser ses propres données.[330] Tout cela en fait une institution compétitive et toujours en recherche d'amélioration vis-à-vis de l'emploi de ses ressources brutes, qui n'hésite pas être innovante dans le développement de son système d'information.

Il semble également que la BBC a su se créer une place en assumant sa position prépondérante en tant que producteur de contenus, même en ce qui concerne la définition de ses identifiants. Elle incarne une stratégie relativement souple à ce niveau-là : un identifiant en intégrant d'autres mais gardant sa spécificité propre.

3.1.3. Archives de France, le service avant la donnée

Contexte et enjeux

Les Archives de France ont, comparativement à la BnF, un historique institutionnel plutôt « récent ». Enfant de la Révolution, l'institution nait en 1794 de la loi du 7 messidor an II, avec pour objectif de conserver les archives des administrations, regrouper les fonds de l'Ancien régime, et prendre en charge les archives saisies comme biens nationaux.[331] Il faudra cependant attendre 1897 pour que la fusion entre le ministère de l'instruction publique des Archives nationales et le bureau des archives du ministère de l'intérieur fasse éclore la direction des archives (prenant le nom en 1936 de la direction des Archives de France)[332]. Cet historique de regroupements successifs jusqu'à former un tout cohérent est assez représentatif de la manière intrinsèque dont sont gérées les archives en France : la prise en compte de fonds distincts et de manières de faire hétérogènes pour constituer un ensemble.

Comme nous l'avons vu, en archives, l'identifiant du monde physique sous-tendait depuis longtemps déjà une logique de cotation.[333] Cette logique de cotation était liée au besoin de ranger, classer un objet, contrairement au monde des bibliothèques où l'idée était plutôt tournée vers le besoin de trouver. En archives, le cadre préexiste, et il fournit un outil qui est un outil de rangement avant d'être un outil de recherche. A partir des années 1940, cette logique se voit contrainte d'évoluer, confrontée à une expansion drastique du volume : les archives produites dans les années 50 à 70 sont déjà plus conséquentes que celles conservées pour l'intégralité des siècles précédents. Ce bouleversement signe la naissance des identifiants opaques, plus maniables. Contrairement là encore aux bibliothèques où le service centralisé le plus important apporte sa norme et pousse les services connexes à s'aligner[334], dans le cas des archives, c'est l'inverse qui se produit. La structure principale accompagne les structures connexes : les structures du réseau

[330] Nous citerons comme exemple la relation symbiotique entre Musicbrainz et BBC News stories, qui s'enrichissent de données l'un l'autre bien que n'ayant pas les mêmes objectifs initialement.

[331] *Historique des archives*. Site Frances Archives.fr [en ligne] Disponible sur https://francearchives.fr/article/37706 [consulté le 07/08/2017]

[332] *Ibid.*

[333] Sauf mention du contraire, l'ensemble des informations qui suivent dans cette partie sont extraites d'un entretien réalisé le 3 mai 2017 avec Romain Wenz, responsable du portail France Archives.

[334] Nous l'avons notamment vu avec les identifiants ARK de la BnF.

sont détentrices des données uniques qui l'intéressent et qu'elle veut mettre en valeur.

Documents manipulés et numérisation

Les archives, contrairement là encore aux bibliothèques, produisent énormément de documents numérisés (plus de 480 millions de documents numérisés en 2015 sur l'ensemble du réseau) [335], mais elles numérisent des séries limitées. Un document d'archive numérique peut ainsi contenir des centaines de feuilles numérisées (états civils, cadastres…). L'identifiant du document est donc parfois à un niveau de granularité très haut par rapport à un identifiant d'image. L'identifiant issu de la numérisation est donc à la fois non pérenne, car il n'est pas une garantie d'accès univoque à un document ; et à la fois pérenne en termes de service : toute la première partie de l'URI est fixe (la racine), et ce qui distingue les entités au sein de celle-ci sont des suffixes non pérennes (sur lesquels l'autorité ne s'engage pas). Par exemple, nous pourrions retrouver le suffixe pérenne « AD45-[cadastre12345] », suivi de « /Folio1, /Folio2, etc. », suffixe modifiable. Cela permet, en cas de problème à la numérisation, de rajouter ou d'enlever des documents sans perturber l'incrémentation de l'attribution.

Nous avons donc des documents qui ont originellement une cote physique, cote attribuée lors du dépôt (processus d'inventaire), et un identifiant de numérisation attribué lors de la création d'une copie numérique. La logique métier employée pour la mise en ligne de ces documents numérisés est donc de dire : s'il y a une cote physique, alors cette entité est susceptible d'être demandée dans un centre de lecture, elle présente un intérêt pour la diffusion en ligne, donc elle mérite d'être identifiée et d'obtenir une URL.

Le projet Frances Archives

Le projet France Archives correspond à un premier vrai pas pour les Archives de France dans l'intégration au web de données. Le précédent site des Archives de France était en effet plus dans une optique de diffusion des contenus sans réellement chercher à développer l'aspect web sémantique. Actuellement, l'ensemble des données présentes sur l'ancien site sont redirigées au moyen du code statut 303 vers des entités du nouveau site, afin de permettre une continuité des contenus. C'est un projet en cours, il y a aujourd'hui environ 30% des métadonnées d'archives du réseau présentes sur le site, ce chiffre devrait atteindre 100% d'ici trois ou quatre années. L'étendue présumée d'un tel projet est évaluée à quelques dizaines de millions de documents qui pourront être référencés en ligne voire accessibles directement pour certains.

Les données publiées sur France Archives ne sont pas hébergées directement sur le site pour la plupart mais constituent un catalogue de métadonnées permettant la redirection vers les sources des structures du réseau national (archives régionales, archives départementales, etc.). Le logiciel de France Archives en interne absorbe les métadonnées des ressources fournies par les unes et les autres afin de constituer une base la plus exhaustive possible. Celle-ci serait à même dans l'idéal de retourner toutes les archives concernant un nom sous forme de notices, qu'elles soient accessibles en ligne par la suite ou non, qu'elles soient disponibles sur le site du

[335] SIAF. *Des Archives en France, l'activité des services d'archives en France*, 2015. [en ligne] Disponible sur https://francearchives.fr/file/99c22fdba4d715d1a5d65a829186c3a5892fddf4/static_9493.pdf [consulté le 08/08/2017]

département de Franche-Comté ou dans des magasins à la Réunion. L'idée est donc de rassembler un maximum de références de documents de manière à faire en sorte que les usagers de la base puissent avoir accès à ces métadonnées et ces descriptions des différents documents. La gestion se fait sur un ensemble de fonds ainsi que sur les éléments à l'intérieur de ces fonds appelés les composants.

En outre, les données présentes sur France Archives sont déjà disponibles sur le site data.culture.communication.gouv.fr qui est le site *open-data* du ministère de la culture et de la communication. Actuellement, la direction des Archives de France fait partie des quelques structures à avoir collaboré et échangé ses données avec celui-ci. Progressivement le site data.culture.gouv sera cependant amené à recueillir les données des autres institutions culturelles : musique, arts du spectacle, bibliothèques, etc.

Modèle de données

Le modèle de données utilisé sur le portail France Archives correspond à la structure EAD (*Encoding Archival Description*), format de description des archives existant depuis les années 90. C'est un format arborescent, permettant de structurer hiérarchiquement par tout-partie et s'appliquant efficacement aux besoins archivistiques en la matière. La structure EAD permet en sus de garantir que le « niveau » du fonds ne va pas changer. En effet, le principe du respect des fonds s'appliquant également au traitement informatique proscrit le morcellement, et donc l'identifiant attribué à un fonds, à plus forte raison s'il est opaque, n'aura pas besoin d'être modifié.

En interne, les données sont converties en RDF. A partir du portail, les données sont récupérables également dans ce format. Il y a un flux XML valide et auto-documenté lors de la navigation qui permet de récupérer l'ensemble des données proposées par le site sur une entité, avec un modèle de données conforme au standard. En outre, de la négociation de contenu est implémentée pour proposer la navigation sur d'autres types de représentation.

L'écueil principal qui doit être évité est le doublonnage : lors d'une mise à jour de la base, d'un import, il y a un risque de création de doublons. Ce risque est d'autant plus important sur une base qui contient déjà six millions de documents. La principale solution trouvée pour permettre d'éviter cela est la gestion de la cote tout au long du traitement de l'objet. Dans le schéma normal, cette cote sert en permanence à vérifier qu'il n'y a pas de doublons, et contribue ainsi à assurer un service simple, propre et surtout efficace.

La double identification des entités

En vue d'être créées, intégrées, gérées et mises à disposition en ligne, les ressources ont plusieurs identifiants tout au long du processus. Le premier identifiant, le plus important et surtout qui suivra le document tout au long de son traitement est la cote d'archive physique. Elle est attribuée lors du dépôt de l'archive et son inventoriage. Cette cote est très normalisée originellement, surtout dans le cas des services départementaux et communaux comme nous l'avons vu. La syntaxe pourra être par exemple « FRAD » pour France Archives Départementales ou FRAM pour les Archives Municipales, suivi du numéro « 005 » pour le département des Hautes-Alpes, un tiret puis un numéro incrémenté. Ces cotes sont primordiales, car elles identifient l'entité de base qui va intéresser l'utilisateur final. C'est sur elles que repose la granularité du fonds.

Mais la mise en ligne et le traitement numérique amènent la nécessité de générer et d'utiliser d'autres moyens d'identification adaptés aux usages. Une même entité pourra donc recevoir plusieurs identifiants reliés les uns aux autres par les métadonnées. L'attribution se fait en général deux fois :

D'une part, en entrée, un premier identifiant opaque est automatiquement attribué en interne à chacune des entités via un petit algorithme, dès que celles-ci sont intégrées au logiciel. Cet algorithme, à l'image des UUID, génère via un système s'apparentant au hashage*[336] un identifiant unique qui a très peu de chances d'être construit une deuxième fois (1 chance sur des milliards). Les données utilisées pour effectuer le hashage correspondent en général à la cote d'archive physique du document de base, qui dans tous les cas sera incluse dans les métadonnées pour justement éviter d'avoir des doublons et permettre l'interopérabilité des systèmes.

D'autre part, en sortie, un identifiant spécifique en URI http pour le web servant à localiser la ressource est mis en place pour chaque notice. Elle ressemble tout à fait au système mis en place par la BnF au niveau des identifiants web ARK. Ceux-ci sont constitués de plusieurs sections, toutes signifiantes :

- Le protocole http://,
- Le nom de domaine indiquant la présence de la notice sur le site Frances archives,
- La chaine de caractères indiquant la présence de la notice dans la section sur les métadonnées d'archives,
- La chaine de caractères indiquant l'instrument de recherche ou le niveau de composant. Au sein de celle-ci se trouvent des séries de lettres qui sont signifiantes dans le cadre du modèle de données interne, par exemple : FA-component, FA-identifier…,
- Le caractère de contrôle (généralement un chiffre).

Ces deux identifiants, internes et externes (gestion/diffusion), sont des coréférences. C'est-à-dire qu'ils coexistent autour d'une même entité, et en délimitent l'utilisation web ou interne. A cette syntaxe dérogent néanmoins deux exceptions :

- Les référentiels de thèmes en SKOS, liés au thésaurus W auquel sont attribués des vrais identifiants ARK, sont hébergés sur le nom de domaine culture.fr/ark/ et non pas sur celui du reste des identifiants locaux, car ils sont liés à d'autres projets thématiques tels que data.bnf.fr, dans un environnement à vocation très pérenne. Ils sont d'ailleurs antérieurs au projet France Archives, et ont été mis en place par les Archives de France en 2011.
- Les pages web éphémères, telles que les pages d'actualités, n'ont pas d'identifiants pérennes car elles ont vocation à être supprimées à terme.

Méthodes d'homogénéisation de la pérennité

Assurer la pérennité des données présentes sur le site est une nécessité pour le projet France Archives. Néanmoins, le fait de travailler également avec les

[336] La mécanique de hashage, ou hash, (à bien distinguer des Hash URI ou dièse que nous avions vus précédemment) est très associée aux systèmes de signature électronique. Il s'agit de récupérer une ressource, la faire absorber par un algorithme qui va coder chaque bit de celle-ci afin de produire un numéro complexe, qui sera complètement différent si le moindre octet était modifié dans le document par la suite. Ce « hash » sera en quelque sorte la carte d'identité unique de cette ressource telle qu'elle est exactement à un instant *t*. Dans notre cas précis, l'algorithme récupère les données qu'il possède déjà sur l'entité afin de créer un identifiant unique selon ce même principe.

partenaires des archives territoriales (départementales, municipales, etc.) amène à déléguer une partie de cet engagement de pérennité à ces structures : comment faire dans le cas où ces partenaires ne proposent pas d'URL pérennes de leur côté? En effet, il n'est pas possible d'exiger le même niveau technique d'un service de dix agents que d'un service à l'échelle d'une métropole. Pour pallier à cet écueil, France Archives utilise le système des requêtes pré-câblées. A partir d'une cote (là encore, la cote physique unique attribuée au dépôt de l'archive) les bases locales sont interrogées, puis la requête est préenregistrée pour fournir lors du clic d'accès à la ressource une URL non-pérenne induisant déjà le résultat cherché. Le logiciel intègre les éléments qui devraient figurer depuis une page source quand on suit un lien vers un site externe. C'est une manière de créer de l'identification pérenne là où il n'y a justement pas d'identifiant pérenne.

Problèmes rencontrés

Les identifiants opaques utilisés peuvent cependant perdre leur pérennité dans certains cas. En effet, dans le cas d'un fonds pour lequel le travail d'inventoriage et d'indexation n'a pas été effectué très en détail, une fois l'identifiant attribué ce n'est plus possible de le détailler ensuite. L'inventaire sommaire d'un fonds, à opposer à l'inventaire complet analytique, apporte une granularité moindre qui maintient des entités comme sous-ensembles alors qu'elles pourraient peut-être constituer en réalité elles-mêmes plusieurs entités.

Prenons comme exemple un fonds personnel constitué des archives d'une personnalité. Celles-ci, lors d'un inventaire sommaire, pourrait être détaillées en plusieurs sous-ensembles, tels que la correspondance, les archives fiscales, les examens de santé, etc. Une fois ces blocs identifiés, ils se voient chacun attribuer un identifiant opaque. Mais quelques temps plus tard, lors d'une révision de l'inventaire, d'autres sous-ensembles sont distingués et remettent donc en cause l'identification pérenne de ces sous-blocs, qui de trois, passent à quatre et ainsi invalident l'identifiant de base. Dans une logique de rangement, trouver un emplacement vide n'est pas difficile, mais dans une logique web, si. Comme nous l'avons vu, la capacité d'un moteur à ne pas retourner d'erreur 404 conditionne son utilisation et son référencement dans les pages de résultats. Le moyen de palier à ce problème est en général l'utilisation des redirections 303, qui sont employées principalement dans le cas d'un besoin de réattribution d'identifiants obsolètes. Cependant, sur France Archives, la séparation des identifiants internes et externes agencés en coréférence (un identifiant de gestion, un URI http) permet au premier d'évoluer en cas de besoin sans impacter le second.

Il n'y a pas non plus une certitude que les identifiants opaques générés soient absolument et totalement uniques, contrairement à d'autres méthodes de création d'identifiants. En effet, malgré le très faible taux de probabilité, le risque existe. Mais un tel cas n'a semble-t-il jamais été encore rencontré aux Archives de France.

Résultats

Nous avons donc une structure qui tente de globaliser son service au moyen d'outils de diffusions interopérables : le site France archives a vocation à enrichir et donner à voir sur un portail commun les ressources mises en ligne par des initiatives individuelles, tout en permettant l'augmentation du trafic et l'utilisation des sources initiales. La logique qui sous-tend cette mise en place est ce que l'on pourrait appeler en termes managériaux de type « *bottom-up* »*, du petit vers le grand acteur. C'est

l'inverse de ce que l'on a pu voir côté bibliothèque avec une logique plutôt « *top-down* », la structure principale centralisant les décisions et influençant les plus petites structures avec l'emploi d'un identifiant global commun.

Les Archives de France choisissent pour l'identification de leurs ressources un système adaptatif souple qui se veut moins normalisé et plus complexe, avec une multiplication des canaux d'identification en fonction des usages : usage physique (cote), usage interne (identification cryptée en hash), diffusion (URI http). Ces différents identifiants, articulés en coréférence, exploitent au maximum les possibilités du numérique sur les utilisations faites des métadonnées. Celles-ci sont la clé de l'interopérabilité des outils : elles permettent d'identifier réellement la ressource en interne, en externe, dans le monde physique comme dans le monde numérique.

Le rôle qui est donc joué par l'institution avec le projet France Archives est celui d'intermédiaire entre :

- d'un côté le web, où sont attribués des identifiants pérennes au sens où les moteurs de recherche et les internautes peuvent les utiliser de manière fiable,
- de l'autre côté un lien vers les applications d'origine où les internautes peuvent retrouver les documents numérisés.

La philosophie des Archives de France est donc extrêmement orientée service. Plus même que la praticité apparente d'un identifiant unique et global, les systèmes, avec cette multiplicité des identifiants, servent avant tout l'utilisation qui en est faite. Le principe est la recherche de la pérennité d'un service qui est construit autour de la donnée : la pérennité de l'accès est valorisée par rapport à la pérennité de la donnée en elle-même. Il s'agit de tisser et de donner accès à une toile qui reste efficace en toutes circonstances pour les utilisateurs, en dépit des consensus généralement attendus en matière de web sémantique et d'identifiants : non-réattribution, pérennité et stabilité de l'identifiant, et limitation de la coréférence au possible.

On retrouve donc ici un contournement habile des questionnements bien propre à l'institution, qui ressemble d'ailleurs beaucoup à ceux rencontrés au niveau de la cotation physique depuis des décennies. Il semblerait que si le passage au numérique n'a pas permis comme on pourrait le penser de débloquer des problématiques anciennes en termes d'identifiant, le web et la mise à disposition en *open-data* apportent finalement les réponses attendues.

3.1.4. Autres initiatives notables

L'Office des publications de l'Union Européenne

Organe interinstitutionnel, l'Office des publications de l'Union Européenne est un éditeur juridique qui « *assure l'édition des publications des institutions des Communautés européennes et de l'Union européenne* »[337]. Parmi celles-ci se trouvent le fameux Journal Officiel de l'Union Européenne, édité en 24 langues. En mars 2012, l'institution met en place la phase de production du Projet Cellar, visant à rendre accessible dans un endroit unique l'ensemble des métadonnées et du contenu numérique produit et géré par l'Office de publication d'une manière

[337] FRANCART, Thomas. *Etude de cas, Office des Publications de l'Union Européenne*. [en ligne] Disponible sur http://www.sparna.fr/reference/office-des-publications-de-lunion-europeenne/ [consulté le 08/08/2017]

harmonisée et standardisée[338]. Les objectifs annoncés sont pluriels : permettre au citoyen un accès plus facile et fiable aux textes de loi, encourager et faciliter la réutilisation des données par les professionnels, et préserver les contenus tout en pérennisant ses accès. Il s'agit donc d'un projet de base de données *open-data* et *open-access* globale à vocation internationale, qui concerne des documents uniques et relativement importants.

En pratique, il y a donc un fort besoin en termes de négociation de contenu, mais également en termes de négociation de langage. Pour le premier, le choix a été fait d'utiliser un serveur http qui leur est propre. Pour le second, si le langage demandé par le *header* du navigateur n'est pas disponible, le serveur retourne une réponse « aucune variable acceptable », ou bien « choix multiple », auquel cas Cellar utilise son propre logiciel pour retourner spécifiquement une représentation[339]. Pour ce projet de base de données en ligne, le modèle de données est construit sur les principes FRBR, et les métadonnées sont représentées en RDF, structurées en DC avec des ontologies en SKOS et OWL. Le langage de requête utilisé est SPARQL. Tous les textes de loi sont publiés avec leur identifiant ELI afin d'être interopérables éventuellement avec d'autres bases.

Chaque version et chaque format a son URI et peut être adressé individuellement. Les identifiants (de diffusion en tout cas) sont signifiants. Lors de la conception de l'URI http, l'équipe a choisi de ne pas inclure le nom Cellar dans le nom de domaine. Le nom employé est publications.europa.eu, plus stable et conservant la substantifique moelle de la mission première du projet. Cela permettrait notamment d'éviter que l'ensemble de l'implémentation technique soit bouleversée si l'Office venait à changer de nom (ce qui est, du reste, très fréquent dans les administrations), le projet à évoluer, etc. La syntaxe des URI est la suivante[340] :

http://publications.europa.eu/(type)/(sous-type)/(identifiant).

Les trois types distingués sont la ressource, l'ontologie, l'application web. L'identification des journaux Officiels est faite grâce au type ressource, puis le reste de l'identifiant est construit avec l'année de publication, le numéro d'édition dans l'année, et enfin la langue.

Par exemple :

http//publications.europa.eu.ressource/oj/JOP_1954_004_R.FRA [341].

Nous voyons ici qu'il s'agit d'une mise en commun de ressources servant à harmoniser à la fois les modes d'accès mais aussi la disponibilité internationale des contenus (en termes de langue, en termes de formats, etc.). Les identifiants signifiants sont pensés de manière simple et efficace afin d'être manipulables au maximum par les usagers. Cela rejoint totalement une volonté de décomplexifier des textes de loi afin de les rendre plus abordables par le grand public, favorisant en parallèle leurs utilisations.

---

[338] SCHMITZ, Peter. Common Access to EU Information based on semantic technology. *The Multilingual Web – The Way Ahead*. W3C Workshop, 2012. [en ligne] Disponible sur https://www.w3.org/International/multilingualweb/luxembourg/slides/41-schmitz.pdf [consulté le 08/08/2017]

[339] ARCHER, Phil, GOEDERTIER, Stijn, LOUTAS, Nikolaos. *Op. Cit.*

[340] *Ibid.*

[341] *Ibid.*

Le secteur public du Royaume-Uni

La mise en ligne en 2009 de data.gov.uk, l'équivalent anglais des projets nationaux d'*open-data* insufflés initialement par Barack Obama aux USA, a permis une introspection dans les administrations sur le thème de la gestion des identifiants utilisés pour leur production de données numériques. Cela a notamment donné lieu à la rédaction d'une note à vocation d'homogénéisation des pratiques, intitulée *Designing URI sets for the UK Public Sector*[342], qui concentre les recommandations liées à la construction et l'attribution d'identifiants en amont de la production de données.

Les départements et agences du gouvernement anglais conservent une liste de tous les types d'entités dont ils sont responsables : écoles, routes, lois, lieux, projets, évènements… L'idée est d'associer individuellement à ces « objets » un identifiant de référence afin de faciliter les échanges internes et la communication. Cet alignement autour d'une sorte de vocabulaire contrôlé commun, nommé VoiD : *Vocabulary of interlinked Datasets*[343], se fait donc autour de ce qu'ils nomment « la donnée de référence », qui correspond à une signification partagée et connue par tous les services et agences associée à cette entité, et qui possède un identifiant unique. Cette idée est basée sur le principe que l'homogénéité est la base de l'interopérabilité.

Le problème demeure : qui attribue l'identification de la donnée de référence ? Le choix a été fait de désigner des départements majeurs en fonction de la correspondance entre leur domaine d'exercice et le type de donnée, afin de leur céder la charge de l'attribution. Les autres départements, services, agences sont ensuite simple utilisateurs. Bien que la provenance soit une information importante, elle n'est pas contenue dans l'URI. Le domaine utilisé pour l'hébergement est très large : data.gov.uk, qui correspond à tous les ensembles d'URI. A celui-ci s'ajoutent des sous-domaines par secteurs, à différencier des départements actuels qui peuvent évoluer (par exemple, education.data.gov.uk).

Ces concepts identifiés sont néanmoins répartis dans des typologies qui sont intéressantes du point de vue de nos questionnements sur la « crise identitaire » (qu'identifie-t-on ?). L'équipe définit 5 grands concepts pouvant être identifiés : les URI Identifiant, liés à des objets réels, concepts, etc. ; les URI document ; les URI liste ; les URI set ; et les URI ontologie. En outre, en termes de syntaxe, les recommandations déterminent deux types de construction d'URI:

- Soit par concept/référence : on identifie un concept (par exemple, une route), en casse basse, les mots séparés par des tirets, au singulier, et ensuite on identifie une instance unique de ce concept (par exemple : http://data.gov.uk/route/RN5),
- Soit par type/format : on identifie un type d'URI par un acronyme (par exemple, pour les URI document on met « doc ») puis on identifie le format .doc, .rdf, .pdf (exemple : http://data.gov.uk/doc/RN5.rdf).

Tous les URI construits sont déréférençables vers un URI document et au moins une représentation de ce document. L'attributaire de l'URI se doit en outre d'implémenter les moyens de rechercher un URI identifiant et d'être redirigé vers un URI document (méthode de l'URI 303), de découvrir les différentes URI de

[342] DAVIDSON, Paul, CIO Sedgemoor, District Council. *Op. Cit.*

[343] Le terme « *void* » en anglais désigne « le néant ». Espérons juste que cela n'est pas symptomatique de l'état d'avancement actuel des choses…

représentation mises à sa disposition, ainsi que de récupérer la représentation la plus appropriée. Pour cela, les auteurs proposent le développement d'un serveur de requête central permettant de résoudre toutes les URI réutilisables. De plus, l'attributaire doit également développer un système d'alerte lors de changements pour les utilisateurs.

Des métadonnées devront aussi être associées lors de l'attribution : la définition des concepts via une ontologie, les relations avec les autres URI, la provenance, le statut officiel, la fiabilité (plus ou moins suivant les cas), la complétude, la régularité des mises à jour, les droits liés, les objectifs de pérennité cités, le public cible, et la description de tous les formats de fichiers qui seront disponibles dans l'ensemble. On retrouve quelques prérequis des identifiants ARK, qui ici sont imaginés de manière beaucoup plus personnalisée et approfondie.

Le secteur public du Royaume-Uni témoigne avec ce projet d'une volonté de profiter des avantages apportées par la technologie du web de données, au-delà du simple fait de publier en *open-data* leurs contenus numériques. Cette démarche fait la démonstration que la réorganisation nécessaire à la publication peut profiter en interne à la structure et augmenter l'interopérabilité et la communication entre les différents services/agences/départements, souvent très cloisonnés.

La Bibliothèque nationale d'Australie

L'Australie est un gouvernement fédéral qui, d'après Diana Dack, consultante sur la question, fait preuve d'une « *méfiance quasi-traditionnelle pour la centralisation* ».[344] A l'aube de l'année 2001, la bibliothèque nationale d'Australie (*National Library of Australia*, NLA) souhaite s'intégrer dans le web de données et surtout mettre en place une meilleure gestion des identifiants de ses ressources. Cependant, à l'époque aucun système d'identifiant pérenne ne convenait complètement aux objectifs de la NLA : l'idée était globalement d'avoir un système nationalement étendu dans lequel la bibliothèque aurait un rôle de coordinateur mais ne serait pas l'élément central. Elle gérerait le système d'attribution en enregistrant les organisations en tant qu'autorités nommantes, établirait un service de déréférencement redirigeant vers les sites des différents partenaires pour une résolution locale, et permettrait l'accès à une résolution centralisée aux structures qui ne peuvent en disposer. De plus, elle proposerait le développement d'un service d'assistance pour aider les organismes adhérant avec un suivi si nécessaire. Ceux qui souhaiteraient s'associer au projet auraient quant à eux l'entière responsabilité de l'attribution des identifiants, tout en respectant un critère globalisant : intégrer une identification de leur structure en début d'identifiant[345]. Pour résumer, la NLA souhaitait qu'une flexibilité et une autonomie maximale soit garantie afin de permettre au maximum de structures de participer au projet et de s'intégrer dans le système.

Les recherches menées afin de trouver un identifiant adéquat ont permis d'envisager 6 possibilités[346] :

---

[344] The National Library of Australia, *Persistent Identification Systems, Report on a consultancy conducted by Diana Dack for the NLA*. 2001. Disponible sur: http://www.imaginar.org/taller/dppd/DPPD/105%20pp%20Persistent.pdf [consulté le 12/04/2017]

[345] *Ibid.*

[346] *Ibid.*

- Utiliser des redirections 303 standard. Il n'y a cependant pas d'engagement de pérennité, la résolution multiple n'est pas possible et il n'y a pas d'outils de management. Tout cela rend la maintenance compliquée,
- Continuer d'utiliser un système de redirection tel que PURL utilisé jusqu'à présent à défaut. Mais celui-ci est dépendant d'un protocole, et de plus il ne supporte pas non plus les redirections multiples,
- Implémenter les URN :NBN. Mais c'est un système peu utilisé, il n'y a pas de service résolution, et ce n'est pas accessible sauf par serveur proxy,
- Implémenter un serveur Handle. Mais c'est plus complexe à mettre en place, ce n'est pas utilisable via les navigateurs standard sans plug-in,
- Adopter le système DOI. Mais c'est relativement couteux, ce n'était pas non plus encore utilisable via les navigateurs standard,
- Adopter le système ARK. ARK semblait une solution prometteuse, mais elle venait tout juste d'être proposée en mars 2001 et donc ne constituait pas une alternative assez mûre malgré les avantages inédits qu'elle présentait.

A la fin de son rapport de 2001, Diana Dack conseille donc de patienter pour l'adoption d'un système et d'en reconsidérer les options dès 2002, afin de voir comment chaque solution a évolué et déterminer la possibilité la plus adaptée au cas de la NLA.

En 2017, la NLA a ouvert deux gros sites de mise à disposition de contenu : Pandora, comportant la collection des archives du web, et Trove, « *une collection de choses merveilleuses* »[347] spécialisée sur les documents numériques (numérisation, journaux, contenus audiovisuel, etc.) Lors de la manipulation de cette base, il semblerait que la NLA ait finalement fait le choix de développer ses propres identifiants pérennes, structurés en URI http avec une centralisation sur l'autorité nommante désignée sous le nom signifiant de « nla ». Par exemple, l'identifiant de l'auteur Rudyard Kipling est le suivant : http://nla.gov.au/nla.party-892127. « *En citant cet identifiant, vous retrouverez tous les documents liés à cette personne. Vous pouvez également relier cette entité avec votre propre service*. »[348]

Cette identification sert aussi bien pour les personnages célèbres que pour les chercheurs, qui peuvent également recevoir leur identifiant unique NLA tout comme chaque étudiant dans l'enseignement supérieur sorti d'une université australienne[349]. Les livres, photographies et autres contenus ont également leur identifiant unique propre, par édition et classées dans des espaces de noms tels que /work/ (œuvres), /version/, ou encore .party pour les personnes ou .arc pour les archives web. La NLA recommande explicitement l'utilisation de ses identifiants pour interopérer sur une base d'identification pérenne[350].

On voit bien ici que la NLA a cédé à la tentation de devenir son propre centre de référence malgré les objectifs posés en 2001.

---

[347] CAMPBELL, Debbie. *Reading locally, learning globally : creating universal experience – a national library view*. NLA, 2010. [en ligne] Disponible sur http://www.nla.gov.au/content/reading-locally-learning-globally-creating-a-universal-experience-a-national-library-view [consulté le 10/08/2017]

[348] *Ibid.*

[349] Ibid.

[350] HOLLEY, Rose. *Resource sharing in Australia : 'Find' and 'get' in trove – making 'getting' better*. NLA, 2011. [en ligne] Disponible sur https://www.nla.gov.au/content/resource-sharing-in-australia-find-and-get-in-trove-making-getting-better [consulté le 10/08/2017]

Conclusion de la partie 3.1

Ces études de cas démontrent que les institutions et les entreprises, en fonction de leurs aspirations, se donnent plus ou moins les moyens de se doter d'une identification opérationnelle. L'implémentation même de cette identification trahit les objectifs (avoués ou cachés) de la structure : ils sont le reflet de l'ambition qu'elle a pour ses données. La BnF souhaite faire de ses données une référence pour tous, la BBC souhaite enrichir son patrimoine en conservant une place prépondérante, les Archives de France souhaitent attirer les utilisateurs et modifier l'image que ceux-ci ont de leurs documents. Les autres projets que nous avons évoqués un peu moins en détail montrent également les volontés des organismes au travers de leur identification : mettre en avant leur accessibilité, se restructurer en profondeur, devenir une référence nationale, etc.

Le système d'identification mis en place possède donc une valeur particulière, il contient des informations en filigrane sur les objectifs des structures. Nous allons à présent voir comment lui faire dire ce que l'on souhaite, et ce à travers deux sous-parties dédiées au re-questionnement des pratiques.

3.2. L'IDENTIFICATION : CE QU'IL FAUT EN DEDUIRE

3.2.1. Le nerf de la guerre

> « Tous les problèmes en informatique peuvent être résolus par une solution de contournement, mais en général ça créé un nouveau problème. »

C'est la citation de David John Wheeler que l'on peut lire en introduction du livre *Linking enterprise data* de David Wood *et al*.[351] En effet, comment ne pas constater après tout ce que l'on vient d'aborder, que la transposition des pratiques dans le domaine numérique a créé de nouveaux problèmes : impossibilité de faire des distinctions claires par support et par genre, complication de la gestion du *versioning* et de la granularité… Même la fameuse question « qui suis-je, où vais-je ?[352] » trouve son incarnation la plus cruelle ici, au sein de la communauté du web de données.

Les enjeux et objectifs d'une institution culturelle sont en adéquation avec les principes et les outils apportés par le web de données, ils sont relativement complémentaires. Mais l'entreprise, elle, a-t-elle vraiment besoin de développer ce genre de projet web (*open data*, *open access*, données liées) ? Les objectifs semblent être multiples : l'entreprise souhaite améliorer ses processus, elle souhaite soutenir le travail des collaborateurs, elle souhaite prendre en compte de nouvelles exigences dans son environnement. Ces projets nécessitent une implication forte sur le long terme avec une équipe dédiée. De plus, les principes fondamentaux du web de données sont en principe déjà présents sur les intranets, même si les techniques en elles-mêmes ne le sont pas.[353]

En termes d'identification, une chose est sûre : il n'y a pas de solution qui convienne à tous les usages ni à tous les organismes, nous l'avons vu en fin de partie 2. Par exemple, beaucoup de projets ont été repoussés à cause des problèmes de lien entre un objet physique, ses métadonnées, la base de données et la représentation numérique de l'objet. Si créer des identifiants uniques (par exemple lors d'une numérisation) est relativement facile, faire en sorte que tous soient liés l'est moins[354]. D'autant plus que le faire manuellement est extrêmement chronophage. Le cœur du problème est donc situé là : comment se comprendre, que ce soit en interne, en externe, entre Homme-machine ou entre machines? Le postulat que l'on pourrait poser à première vue serait de dire : il faut que tout le monde utilise les mêmes outils. Nous l'avons vu en seconde partie, c'est impossible.

Donc, si tous les systèmes n'arrivent pas à s'homogénéiser, il faudra bien au bout du compte créer un langage de communication, échanger et tenter de se comprendre par un moyen ou un autre. En bref, interopérer. Or, quelles sont les possibilités en termes d'identification pour rendre des systèmes interopérables ? Emmanuelle Bermès distingue deux méthodes[355] :

---

[351] WOOD, David. *Linking Enterprise Data*. . Springer, 2010.

[352] « *Dans quel état j'erre ? Où cours-je ?* » citation de Bruno Masure.

[353] DALBIN, Sylvie, BERMES, Emmanuelle, ISAAC, Antoine, WENZ, Romain, NICOLAS, Yann, MERABTI, Tayeb, ANGJELI, Anila, FRANCART, Thomas, ROZAT, Lise, VANDENBUSSCHE, Pierre-Yves, VATANT, Bernard, RAIMOND, Yves, COTTE, Dominique. Approches documentaires : priorité aux contenus. *Documentaliste-Sciences de l'Information*. ADBS, 2011. N° 48. p42-59

[354] VAN HOOLAND, Seth. VERBORGH, Ruben. *Op. Cit.*

[355] BERMÈS, Emmanuelle, ISAAC, Antoine, POUPEAU, Gautier. *Op. Cit.*

- Le *Hub and Spoke** (roue et essieu) : comme pour l'alignement d'ontologies, il correspond à la méthode de rassembler des informations dans un référentiel pour normer les termes et les URI, par exemple en créant des méta-référentiels tels qu'un thésaurus, un vocabulaire contrôlé, une liste d'autorité… Un référentiel qui serait au-dessus des autres et qui permettrait de les lier ensemble.
- Le *Follow your nose**, ou navigation intuitive, soit la réutilisation des jeux de données existants.

> « Si les responsables des jeux de données font le choix de réutiliser des URI existantes au lieu de générer leurs propres URIS locales […] on passe directement du nouveau jeu de données ainsi oublié à celui dont on réutilise les URI. Si une bibliothèque décide, plutôt que de générer des URI pour les auteurs de ses ouvrages, de réutiliser des URI existantes, celles de DBpedia par exemple, il devient possible de naviguer directement non seulement de cette bibliothèque à DBpedia, mais aussi directement vers le jeu de données d'une autre institution, fonds d'archives, musée, qui aurait fait le même choix. »[356]

C'est donc hautement intéressant. Pour paraphraser Seth Van Hooland et Ruben Verborgh, malheureusement, les jeux de données, c'est un peu comme les sous-vêtements, tout le monde est d'accord pour dire que c'est merveilleux mais personne ne veut utiliser ceux des autres.[357]

3.2.2. Question de points de vue…

Nous avons vu en partie 1.2 que les deux approches, Cartésienne et Piercienne, nous donnent une première vision de ce que l'on considère comme la problématique de la provenance : à savoir qui est compétent pour déterminer ce qu'un signe veut dire ? L'approche Cartésienne nous dirigeait vers une idée que l'on peut facilement transposer au web de données : on identifie en trouvant une autorité de référence qui indique ce que chaque signe veut dire.[358]

Catherine Legg, dans son article[359], prend comme cas d'analyse 3 systèmes différents issus de l'approche Cartésienne, afin de démontrer que cette approche n'est pas, en général, couronnée de succès :

- RDFS, le schéma de données de RDF. Il est trop simple logiquement parlant pour vraiment s'adapter à tout ce qu'il y a à définir dans le monde. Il définit les termes de manière typiquement Cartésienne, de manière « autoritaire ». Même si quelqu'un pouvait ajouter des nouvelles classes et les remplir avec de nouvelles instances, il ne pourrait pas dire plus avec, cela manque de souplesse, il est trop opaque. Par exemple, une seule classe primaire rdfs :Class est utilisée pour représenter une infinité de niveaux de classes différents[360].
- OWL, qui reprend les composants de base de RDFS. Il est très peu utilisé en dehors du contexte académique. Il y a notamment des problèmes d'alignement

---

[356] *Ibid*. p46

[357] VAN HOOLAND, Seth. VERBORGH, Ruben. *Op. Cit.*

[358] LEGG, Catherine. *Op. Cit.*

[359] *Ibid.*

[360] PAN, Jeff Z., HORROCKS, Ian. *Metamodeling Architecture of Web Ontology Languages.* University of Manchester, 2001. [en ligne] Disponible sur http://www.cs.man.ac.uk/~horrocks/Publications/download/2001/rdfsfa.pdf [consulté le 31/08/2017]

entre RDF et OWL, car il est maladroit et trop « verbeux ». La structure arborescente ne s'adapte que très peu aux applications réelles car il ne permet pas l'ajout de type de données personnalisées.

- Enfin, CYC, l'interface en langage naturel. Très ambitieux en termes de taille, il possède son propre langage interne à CycL. Il fonctionne à l'aide d'assertions axiomatiques et de règles, mais le système des catégories le rend très complexe.

Nous le constatons, il est vrai que ces outils sont assez peu utilisés, à cause le plus souvent de leur souplesse toute relative, ou encore de leur structure. Seules les grosses institutions peuvent se permettre de les mettre en place, car cela demande d'importants moyens. Le modèle Cartésien est issu de cette idée héritée de la vision de l'intelligence artificielle des années 50, une encyclopédie dans la tête d'un robot[361], c'est-à-dire une vérité unique et absolue émanant d'un point de vue qui serait parfait.

A contrario, l'approche Piercienne, elle, produit des outils que l'on croise au quotidien et qui fleurissent dans le « grand web ». Ils sont parfois référés par le W3C comme des outils de « *semantic web lower-case* »[362], par opposition aux efforts normalisants officiels du consortium. Parmi ces outils, les sites web collaboratifs : quelles que soient leur langue les usagers discutent et définissent des concepts (c'est le cas de Wikipédia, Musicbrainz…). La qualité et la vérité émergent d'une quantité d'opinions et d'utilisations. C'est également le cas des folksonomies* (*tagging*), totalement libres. Certains sites se sont développés pour organiser les pratiques de tag, Delicious ou encore Flickr[363]. L'idée est que « *tout comme nous parlons d'un corps en mouvement et non d'un mouvement dans le corps, nous sommes dans la pensée, ce n'est pas la pensée qui est en nous* »[364] : le web de données évolue seul et se construit par les utilisations qui en sont faites. C'est le cas de Google :

> « Le génie de Google a été de se rendre compte qu'ils n'avaient pas besoin de payer des gens pour inspecter et noter les sites web, car cette donnée existait déjà dans la manière dont ils étaient tous liés entre eux. »[365]

Mais nous pourrions opposer un argument, celui de Cory Doctorow, qui explique que les gens mentent, sont feignants, sont stupides…[366] On ne peut raisonnablement pas compter sur leur travail pour la définition des métadonnées se rapportant contenus. De même, Mirna Willer et Gordon Dunsire considèrent les « amateurs » de l'information comme des producteurs de métadonnées auxquels il ne faut surtout pas faire confiance aveuglément. D'où l'intérêt de connaitre la provenance les métadonnées que l'on rencontre sur le web.

Pour autant, tout le principe de la réponse par l'approche Piercienne tient dans la quantité d'opinions : l'impact de l'individualité faiblit drastiquement avec la quantité. De plus, ces méthodes sont très peu couteuses, elles sont simples, évolutives, et leur succès parle de lui-même : s'il n'y avait personne pour les maintenir et les encourager, elles perdureraient quand même. Wikipédia

[361] *Ibid.*

[362] Web sémantique de « casse basse », ou de basse catégorie, référant à un web officieux plus qu'officiel, maintenu par des non-professionnels.

[363] LEGG, Catherine. *Op. Cit.*

[364] Citation de Pierce, dans LEGG, Catherine. *Op. Cit.*

[365] LEGG, Catherine. *Op. Cit.*

[366] DOCTOROW, Cory. *Op. Cit.*

typiquement s'oppose à des silos de savoirs manuellement figés[367]. David Milne, Olena Medelyan et Ian H. Witten font l'expérience en 2006[368] de comparer la qualité d'un thésaurus développé en miroir à Wikipédia à un thésaurus établit spécifiquement par des professionnels. Ce premier thésaurus se révèle beaucoup plus rapide à construire et complet que pour les mêmes sujets en OWL. Par exemple, au sujet d'une requête faite à propos d'un arbre spécifique de Nouvelle-Zélande, Wikipédia couvre deux fois plus de concepts qu'Agrovoc, l'ontologie manuellement créée sur les sujets spécifiques de l'agriculture[369].

Les approches de Corey Doctorow, Mirna Willer et Gordon Dunsire ne sont pas si incompatibles avec celle de Catherine Legg : les vocabulaires contrôlés peuvent tout à fait enrichir les outils Pierciens et en retour ceux-ci peuvent profiter aux outils classiques. David Wood confirme cette idée en déclarant qu'il y a au moins deux sources de métadonnées de très haute qualité qui sont fiables : les ontologistes professionnels et les « *hive mind** » des réseaux sociaux. « *Il apparait que finalement nous n'avons pas besoin de nous connaitre nous-même, puisqu'on se connait tous un peu entre nous*. »[370] L'intelligence du web n'est pas à construire, elle est à exploiter, car tout est déjà là[371]. Mais le point important à retenir est qu'il ne faut pas dénigrer les approches folksonomiques. En effet, Mirna Willer et Gordon Dunsire le reconnaissent d'ailleurs dans leur ouvrage. En ce qui concerne les bibliothèques en tout cas, il y a déjà un changement de paradigme[372] :

- Les relations entre les ressources bibliographiques sont plus importantes que les descriptions de ressources individuelles,
- Il est accepté que les notices ne seront jamais totalement complètes,
- L'humain peut se contenter d'une notice et de métadonnées imparfaites car il reconnait la variation et le contexte,
- L'univers bibliographique est intégré dans l'univers plus large du langage humain,
- De la métadonnée de faible qualité ouverte est toujours préférable à de la métadonnée de très bonne qualité non diffusée ou non accessible,
- Le rôle du catalogueur est celui d'un « berger » des métadonnées, et non celui d'un gardien (donc un coordinateur, pas un censeur),
- ….Est-il besoin de continuer ?

3.2.3. Autour de l'objet identifié

Bergman nous dit que « l'espoir utopique » placé dans le web sémantique qui consiste à régler tous les problèmes de communication entre machines est « *idiot et naïf* ». « *Nous ne prenons pas pour argent comptant tout ce que dit un humain, il ne faut donc pas se dire que toutes les transmissions entre machines sont strictement fiables*. » [373]. Le contexte, nous l'avons vu, est primordial. La langue est contextuelle par essence, ce qui n'est pas le cas du fonctionnement des machines qui prennent

[367] LEGG, Catherine. *Op. Cit.*

[368] MILNE, David, MEDELYAN, Olena, WITTEN, Ian H. *Mining Domain-Specific Thesauri from Wikipédia : a case study*. Department of Computer Science, University of Waikato. Hamilton, 2006. [en ligne] Disponible sur http://www.cs.waikato.ac.nz/~ihw/papers/06-DM-OM-IHW-wikipedia_vs_agrovoc.pdf [consulté le 25/08/2017]

[369] *Ibid.*

[370] Wood p149

[371] LEGG, Catherine. *Op. Cit.*

[372] WILLER, Mirna, DUNSIRE, Gordon. *Op. Cit.*

[373] BERGMAN, Mike. *Op. Cit.*

tout « au pied de la lettre ». Le problème, c'est que de ce fait, les mots, les noms, les références, les identités et les sens ne sont pas absolus. Penser que ça pourra un jour être le cas dans la communication avec les ordinateurs est tout aussi naïf. Afin d'affecter dans le bon sens la communication entre ordinateurs nous devons pouvoir observer, tirer des leçons et appliquer les meilleures règles issues des interactions humaines. Pour cela, Bergman propose trois idées [374]:

- comprendre et faire confiance de manière systématique à des sources repérées comme fiables,
- interpréter et contextualiser pour juger si des sources sont valides et appropriées,
- et enfin effectuer des tests de validation pour garantir la valeur des messages reçus.

En effet, le principe de base du web de données est la capacité à formuler des affirmations, des déclarations construites comme des phrases sujet-verbe-objet. Pour que ces phrases soient exploitables il faut être sûr que les éléments sont non ambigus. Mais alors, comment distinguer les fines différences qui peuvent exister entre un concept, son nom, sa localisation, sa représentation ? Ce sont tous des éléments très proches mais impliquant au niveau sémantique des idées complètement différentes. Cette dissociation entre une URI et son contenu est nommée « Principe de l'Opacité* »[375]. David Booth propose deux méthodes pour faire la distinction [376]:

- Utiliser différents noms pour les distinguer. Il faudrait un nom pour l'objet, un nom qui désigne le nom, un nom qui désigne le nom du nom… (Fred, name :Fred, name : name : Fred…) cela peut être infini.
- Utiliser différents contextes en employant le même nom. Syntaxiquement, c'est possible, c'est le cas par exemple des qualifiants ARK : un tronc commun référant à une entité unique identifiée, puis des ajouts variables en tant que suffixe pour distinguer des éléments de granularité, des représentations, etc. Des conventions syntaxiques peuvent être utilisées, mais également d'autres types de conventions (par exemple un environnement technique, métadonnées, etc.)

En pratique, ces deux méthodes induisent tout de même un changement de l'URL, ce qui potentiellement revient au même.[377]

Quant à savoir ce qu'il est important de publier, au-delà même de la donnée, c'est une question primordiale qui doit se trouver dans les prémices du développement des projets. Nous l'avons vu, les Archives de France déclarent que ce qui est coté à la base mérite de ce fait d'être publié. Si un objet a été identifié, c'est-à-dire préalablement distingué, sélectionné, individualisé, c'est qu'il est susceptible d'intéresser quelqu'un. Pour Norman Paskin et Godfrey Rust, faire la distinction entre ce qui doit être identifié, au sens philosophique du terme, et ce qui ne doit pas l'être, c'est « *séparer les choses qui sont pareilles entre elles et les choses qui diffèrent pour une quelconque raison.* »[378] Aucune discrimination donc

[374] *Ibid.*

[375] JACOBS, Ian, WALSH, Norman. *Architecture of the World Wide Web, Volume One*. W3C Recommendation, 2004. [en ligne] Disponible sur https://www.w3.org/TR/webarch/ [consulté le 09/08/2017]

[376] BOOTH, David. *Op. Cit.*

[377] *Ibid.*

[378] PASKIN, Norman, RUST, Godfrey. *Op. Cit.*

ici à part celle d'être unitairement distinguable. Cela fait référence à la loi de l'identité de Liebniz, que nous avions évoquée en première partie de ce mémoire[379].

Un autre point de vue sur la question émane de l'ontologiste John Sowa, qui considère que comme la loi physique fait que deux objets ne peuvent pas être dans le même volume physique (au sens de l'espace occupé) au même moment, l'espace-temps peut être utilisé comme identifiant[380]. Cependant cette vision, bien qu'elle soit tout à fait pertinente, rentre totalement en conflit avec les recommandations de la communauté web de données sur la non-introduction de données techniques dans l'identifiant. Est-ce là une occasion de reconsidérer la question ?

Le *vade mecum* du ministère de la culture détermine quant à lui deux critères pour savoir si un document doit être identifié[381] : est-ce que la pérennité du référent sera amenée à être intéressante ? Par exemple, choisir la version que l'on va garder d'un document, ne pas identifier les autres… Est-ce que le document est réutilisable ? Surtout si celui-ci sera amené à être réemployé souvent, citation, lien avec d'autres entités, etc. A ceci nous ajouterons simplement un bémol : est-ce que des Vincent Van Gogh des temps modernes dont les « croutes » ne seraient pas jugées dignes d'être identifiées pourraient ainsi manquer aux générations futures ?[382] Et puis, comment déterminer ce qui « mérite » d'être pérenne et ce qui ne le mérite pas ?

3.2.4. La problématique de la pérennité

La problématique de la pérennité est, avec celle de l'interopérabilité, un des sujets les plus discutés dans le domaine du web de données et plus généralement de l'identification. Selon Sébastien Peyrard, Jean-Philippe Tramoni et John Kunze, les utilisateurs s'attendent à avoir un contenu stable derrière les liens de contenus publiés, mais ils s'attendent à du contenu dynamique derrière les identifiants pérennes de contenu éditorial (une page d'accueil, une base de données…). Les deux situations sont des cas légitimes d'utilisation des identifiants pérennes, même si le second déroge avec les recommandations généralement évoquées en ce qui concerne l'identification. A ce sujet, Margaret Byrnes évoque 4 types de variabilité des contenus[383] :

- Le contenu corrigible : enregistré auparavant qui nécessite d'être corrigé à tout moment,
- Le contenu dynamique : enregistré auparavant remplacé par du contenu continuant à correspondre aux métadonnées de description (par exemple : une page météo réactualisée avec les données du jour qui serait identifiée de manière pérenne),

---

[379] « *Pour parler simplement, dire que deux choses sont identiques n'a pas de sens, et dire qu'une chose est identique à elle-même c'est ne rien dire du tout* ». Ludwig Wittgenstein, à propos de la loi de l'identité de Liebniz.

[380] SOWA, John. Re : [ontolog-forum] OWL and lack of identifiers. Ontolog-Forum, 2007. [en ligne] Disponible sur http://ontolog.cim3.net/forum/ontolog-forum/2007-04/msg00030.html [consulté le 09/08/2017]

[381] Ministère de la Culture et de la Communication. *Op. Cit.*

[382] De toute évidence, ceci n'est qu'une grossière métaphore pour simplement dire que potentiellement, le regard que l'on pourra porter sur ce qui est utile, réutilisable ou pérenne peut éventuellement évoluer avec les années : des données que l'on pourrait croire secondaires pourraient effectivement être mises de côté jusqu'à ce que l'on se rende compte qu'elles auraient été intéressantes. Mais peut-être ces idéaux d'absolu sont totalement utopiques et qu'il faudra effectivement bien faire un tri à un moment donné…

[383] BYRNES, Margaret. *Defining NLM's Commitment to the Permanence of Electronic Information*. 2000. Cette référence n'étant apparemment plus accessible en ligne, c'est de la citation présente dans le document PEYRARD, Sébastien, TRAMONI, Jean-Philippe, KUNZE, John. *Op. Cit.* que nous tirons cette référence.

- Le contenu invariable : qui ne bouge jamais, mais le format d'encodage et les hyperliens peuvent évoluer (par exemple lors d'une migration de format),
- Le contenu en flux : un flux de bits représentant un contenu qui ne bouge pas.

Pour les auteurs de l'ISA, Phil Archer, Stijn Goedertier et Nikolaos Loutas, « *la meilleure garantie de pérennité est l'utilité* »[384]. Une identification pérenne garantit que toutes les ressources sont identifiées, que les URI sont permanentes et stables, qu'elles sont facilement maintenables, uniques, claires, concises, qu'elles sont explicitement liées les unes aux autres, et qu'elles sont intuitives. A ceci s'ajoutent les recommandations classiques : il ne faut pas qu'elles contiennent de mot-clé, ni d'extension de fichier, il faut qu'elles soient systématiquement en casse basse, ne contiennent ni accent ni espaces, remplacent les caractères spéciaux, etc.[385]

Mais pour John Kunze[386], la vision d'un identifiant parfait qui règlerait les problèmes de pérennité est très éloignée de la réalité, et elle est incompatible avec une certaine gestion à terme de la pérennité. Pour lui, la pérennité est purement un problème de « service ». Il peut en effet y avoir différents NMA[387] qui fournissent le même accès au même document (avec un même identifiant) mais qui ne garantissent pas les mêmes garanties en terme de pérennité. La plupart des serveurs de données possèdent déjà des fonctionnalités de redirection (qui sont disponibles « de série » depuis les années 90). Les NMA qui ne prenaient déjà pas le temps de les tenir à jour pour leurs URL ne le feront pas plus pour des URN ou des Handle. Cette vision est également partagée par Tim Berners-Lee qui l'évoque dès 1998 dans son document *Cool URI's don't change*[388]. La seule manière de juger de la persistance d'un fournisseur serait de voir depuis combien de temps il propose ce service et quelle réputation il a. De ce fait, l'implication de différents acteurs dans l'ensemble du processus complique encore la question[389].

John Kunze évoque en outre un autre problème lié : pourquoi dans ce cas ne faisons-nous pas confiance aux URL comme identifiant pérennes[390] ? Tous les identifiants déréférençables sur le web sont des URL, alors comment sont distingués ceux qui renferment une promesse de pérennité ? L'appréhension que l'on a de l'URL en tant que moyen d'identification pérenne est erronée, car elle est basée sur le fait que l'on considère l'URL comme une simple adresse, une localisation. En réalité, la résolution d'URL implique de nombreux intermédiaires, serveurs proxy, serveurs web, opérations d'aiguillage etc., qui rendent l'URL tout aussi indirect (et possiblement opaque sur la réelle localisation des données) que n'importe quel identifiant issu d'un système d'identification pérenne (PID), URN, Ark, Handle…[391]

Cependant, John Kunze approuve totalement l'affirmation des auteurs de l'ISA (l'utilité fait la pérennité). Il évoque la fâcheuse tendance à régler le problème de la pérennité des ressources en proposant une redirection systématique (systèmes de

---

[384] ARCHER, Phil, GOEDERTIER, Stijn, LOUTAS, Nikolaos. *Op. Cit.*

[385] *Ibid.*

[386] KUNZE, John A. *Towards Electronic Persistence Using ARK Identifiers*. California Digital Library, 2003. [en ligne] Disponible sur https://wiki.ucop.edu/download/attachments/16744455/arkcdl.pdf [consulté le 19/05/2017]

[387] *Name Mapping Authority* : autorité/organisme/système qui donne accès et fournit les métadonnées.

[388] BERNERS-LEE, Tim. *Cool URIs don't change.* W3C, 1998. [en ligne] Disponible sur https://www.w3.org/Provider/Style/URI [consulté le 31/03/2017]

[389] KUNZE, John A. *Op. Cit.*

[390] Cette affirmation est explicitée dans la partie 1.2 de ce mémoire.

[391] *Ibid.*

PURL ou URI 303, par exemple). Celle-ci est inspirée du système DNS, pour laquelle cette solution a eu un succès fulgurant, mais cela induit en erreur car les enjeux ne sont pas du tout les mêmes. Dans le cas de DNS, quand il y a un problème avec une machine et que celle-ci tombe en panne (l'identifiant mène donc à un lien mort), la redirection systématique permet de signaler le souci et entraine une réactivité immédiate. Il s'agit là d'un outil très utilisé qui, typiquement, possède une valeur marchande et implique des enjeux « pratiques ». Dans le cas des ressources identifiées, beaucoup plus nombreuses, de valeur moindre et pas forcément utilisées au quotidien, ce principe ne permettra pas forcément la réactualisation.[392]

La notion d'identifiant est également discutée par John Kunze, pour lui l'identifiant n'est pas une simple chaine de caractères mais l'association entre une chaine de caractères et une entité. Pour que cette association soit actée, elle nécessite d'être explicitée par une trace (que celle-ci soit sous la forme d'un catalogue, de métadonnées…)[393]. La liaison peut se faire physiquement, si par exemple l'entité porte la chaine de caractère sur elle, comme dans le cas des cotes archivistiques.

La problématique de la pérennité est une responsabilité qui doit être prise en compte à l'origine du projet et sous l'angle de la gestion du risque[394] : d'un côté il faut étudier le niveau de maturité d'un système d'identifiants pérenne (par exemple, pour pouvoir estimer le coût que peut prendre la mise à jour constante des URL cassées), et de l'autre estimer le risque que l'information devienne inaccessible si le service n'est pas adéquat.

Conclusion de la partie 3.2

Il semble clair que ces questions d'interopérabilité, de pérennité, d'identité sont loin d'être réglées, mais nous pouvons déjà entrevoir quelques éléments de réponse. Les avancées technologiques se sont mieux adaptées aux besoins à mesure qu'elles se perfectionnaient, et le recul apporté par l'analyse de l'évolution des pratiques ainsi que des retours d'expérience qui ont été correctement documentés porte déjà quelques fruits. Après avoir évoqués certains des points de vue et des opinions intéressantes sur les problématiques intrinsèques au web de données, il est pertinent de regarder plus en détail les recommandations techniques que l'on peut en retirer, qui ne proviennent d'ailleurs parfois pas du tout des mêmes sources.

[392] *Ibid.*

[393] *Ibid.*

[394] DAVIDSON, Paul, CIO Sedgemoor, District Council. *Op. Cit.*

3.3. CONDENSE PROCEDURAL DE BONNES PRATIQUES

3.3.1. Conseil n°1 : Assurer en amont l'interopérabilité et la pérennité

La toute première étape qui constitue tout projet, nous le savons, est l'analyse des besoins. En matière d'identification, cette analyse est primordiale car elle permet de définir tout le cadre d'action future : quels sont les objectifs, quelles sont les capacités de la structure, quels sont les éléments que l'on souhaite identifier ? De quelle base démarrons-nous ?

En premier lieu, l'analyse de l'existant apporte déjà des réponses quant aux possibilités qui s'offrent à nous. La question importante à se poser concerne la réutilisation, comment exploiter au mieux ce que l'on possède déjà et ne pas partir de rien ? Nous l'avons vu, il est plus adéquat de chercher à réutiliser ce qui se fait déjà (tant en terme de référentiel que d'identification ou d'attribution de métadonnées, semblerait-il) [395], afin d'assurer au mieux la première contrainte : l'interopérabilité[396]. Les questions que nous allons devoir nous poser sont :

- Quels sont les identifiants déjà disponibles pour mes ressources (en interne, mais également en externe si celles-ci sont aussi partagées par d'autres)?
- S'il préexiste un système d'identification, toutes les entités sont-elles déjà identifiées ?
- Ces identifiants sont-ils complets ?
- Sont-ils localement uniques ?
- Sont-ils fonctionnels ?

Dans le cas où l'on a des identifiants « métiers », attribués lors du passage ou de la création dans une application (ou lors d'une numérisation, d'un processus spécifique, etc.), il est possible de rendre ces identifiants globalement uniques simplement grâce à l'ajout d'un code producteur identifiant la provenance. C'est d'ailleurs une méthode dont nous constatons l'utilisation par les grands systèmes d'identification majeurs : ARK, DOI, ISBN, ISSN… Ils créent des « paliers » d'identification, avec une incrémentation par zone, allant progressivement dans la granularité : pays, zone, structure, sous-structure, identifiant de l'objet.

Il est également possible d'incorporer les identifiants anciens dans des nouvelles URI, mais dans ce cas, il est vital de les réutiliser sans en changer la sémantique originale (par exemple, un numéro de police d'assurance de voiture ne devrait pas être réutilisé dans un identifiant pour le véhicule en soi)[397].

Dans tous les cas, quelle que soit la stratégie individuelle choisie, que l'on ait fait le choix de recréer de toutes pièces un identifiant, d'implémenter des PID ou de réutiliser les identifiants existants, il est très important de toujours conserver une trace des précédents identifiants utilisés. Et ce, à n'importe quelle étape du projet, que ce soit avant la mise en place d'un système global ou même en cours du projet lors d'une montée de version. Nous nous rappellerons notamment à ce sujet le cas Europeana (cité en partie 1.3) qui a attribué plusieurs fois des identifiants pérennes

[395] PEYRARD, Sébastien, TRAMONI, Jean-Philippe, KUNZE, John. *Op. Cit.*

[396] En effet, nous l'avons vu, l'interopérabilité est ce qui rend possible beaucoup de choses : elle permet au système de s'enrichir d'autres systèmes, elle permet de faciliter l'adoption de notre propre système, elle permet d'avoir une meilleure garantie de pérennité notamment pour les futures migrations, etc.

[397] ARCHER, Phil, GOEDERTIER, Stijn, LOUTAS, Nikolaos. *Op. Cit.*

successivement à des mêmes jeux de données, car la trace de la précédente identification n'avait pas été gardée dans les métadonnées.[398] La rétroactivité permet donc d'éviter des pertes importantes de temps, d'argent et d'énergie.

Si on ne réutilise pas ses anciennes URI, il faut néanmoins faire en sorte que les nouvelles n'aient rien à voir avec les précédentes, afin d'éviter la collision entre elles. Les méthodes de protection des nouveaux identifiants peuvent être simples, comme l'ajout d'un préfixe entre nom de domaine et identifiant (ISBN10 devenant ISBN13). L'homogénéisation des identifiants permettra également d'opérer des fusions et des scissions entre les URI qui désignent des mêmes ressources.[399] En effet, l'implémentation d'identifiants occasionne souvent un grand ménage et une chasse aux doublons.

Quant au cas d'une configuration incitant à l'utilisation de PID, les critères de choix devront être :

- l'indépendance proposée vis-à-vis de la technique : privilégier les systèmes simples et efficaces qui peuvent être facilement remplacés, et mettre en place une gouvernance réfléchie qui se tient en dehors toute implémentation technique,
- la popularité du système dans le domaine choisi : préférer les systèmes d'identifiants qui ont fait leurs preuves et qui se sont imposés dans la communauté à laquelle nous appartenons, c'est aussi une garantie de plus vers l'interopérabilité et la pérennité,
- la liberté d'attribution des identifiants : avoir la possibilité de créer des identifiants d'entité qui correspondent à nos besoins précis et qui peuvent nous être spécifiques tout en ayant une gouvernance orientée vers le global,
- ainsi que les contraintes d'engagement vis-à-vis de la pérennité : s'engager soi-même sur des critères de maintenance stricts et s'assurer que le fournisseur des identifiants suit également ces principes, afin de construire une chaîne valide.

Ces critères sont réellement universels dans le sens où il sera toujours moins intéressant d'être lié à une implémentation technique contraignante que d'être sur une gouvernance globale pensée en retrait de la technique, et ce quels que soient les objectifs de la mise en place d'un système d'identification (qui dépendent entre autres de la taille et des missions de la structure). En outre, l'institution qui maintient ces URI doit être digne de confiance, car les risques sont nombreux, tels que les mises à jour compromettant la cohérence des données, les évolutions du système d'identification, la désertion du système, l'obsolescence technique et structurelle, etc.[400]. Le choix de la résolution d'URI par un tiers peut être intéressant, car il permet d'adresser à la fois le problème de confiance lié aux métadonnées renseignées par l'attributaire, et à la fois celui du coût de la maintenance en interne[401]. De plus, favoriser la diversité des services de distribution et d'attribution d'URI et leur indépendance permet d'apporter une garantie de plus à la pérennité de ces identifiants.[402]

---

[398] *Ibid.*

[399] PEYRARD, Sébastien, TRAMONI, Jean-Philippe, KUNZE, John. *Op. Cit.*

[400] ARCHER, Phil, GOEDERTIER, Stijn, LOUTAS, Nikolaos. *Op. Cit.*

[401] WOOD, David. *Linking Enterprise Data*. Springer, 2010. Chapitre "Reliable and Persistent Identification of Linked Data Elements", p149-177.

[402] ARCHER, Phil, GOEDERTIER, Stijn, LOUTAS, Nikolaos. *Op. Cit.*

> « Toutes les stratégies actuelles de systèmes d'identifiants pérennes requièrent une implication humaine afin de garantir la pérennité à long terme et la résolvabilité. »[403]

Enfin, les moyens humains déployés sont toujours primordiaux, tant dans la mise en place d'une gouvernance que pour l'implémentation, sans oublier la maintenance qui doit être assurée tout au long de la vie du système.

3.3.2. Conseil n°2 : Choisir une structure d'URI adéquate

En termes de construction propre de l'URI, même si le choix est fait de les déployer en interne individuellement, les recommandations semblent systématiquement se tourner vers l'utilisation d'URI http, dont l'une des caractéristiques est de faciliter la manipulation. En effet, l'objectif intrinsèque des l'identifiants est qu'ils soient communiqués. Ils sont la plupart du temps utilisés dans ce seul but, transcendant les domaines d'activité et les catégories d'utilisateurs. Il est donc recommandé de se préparer à des cas d'utilisation larges et variés, même si au départ il ne semble pas que cela soit réellement nécessaire.[404] Et si l'information n'a pas d'URL, elle n'existe pas sur le web. Le déréférencement via un navigateur est de surcroit extrêmement pratique et intuitif pour notre société immergée dans l'hypertextuel. Ce n'est évidemment pas le choix idéal pour tous les cas, notamment sur des organismes à la politique de confidentialité spécifique, qui ne souhaitent pas publier sur le web mais pouvoir lier leurs données en interne, ou utiliser les technologies pour leur intranet par exemple, etc.

Selon *l'EC Informal Working Group on Persistent URIs*, le groupe de travail dédié de l'ISA, les URI devraient toujours avoir une structure telle que [405]:

(racine URI : scheme+ autorité)/(chemin ou *path*)/(type ou provenance)/(chaine de caractère)/(options)

Les spécificités fonctionnelles de la syntaxe même d'une URI définies par Tim Berners-Lee sont du même ordre[406] :

(Scheme)/(autorité)/(chemin ou *path*)/(requête)/(fragment)

Nous constatons que seuls la racine et le chemin de la ressource sont nécessaires à toutes les URI, même si celles-ci ne sont pas des URI http. Par exemple, certains URN rentrent également dans ce cadre bien qu'ils ne présentent pas d'autorité ni de fragment[407]:

urn :(racine) exemple : animal : toucan : bec(chemin)

Le chemin peut présenter une organisation hiérarchique pour spécifier le périmètre de l'URI. C'est également un moyen de contextualiser ou d'apporter des éléments lisibles et compréhensibles pour la manipulation de la ressource, par exemple. Les options sont quant à elles des suffixes modulables permettant d'apporter différents services, elles ne comportent généralement pas de garantie de pérennité contrairement au reste de l'identifiant.

---

[403] DAVIDSON, Paul, CIO Sedgemoor, District Council. *Op. Cit.*

[404] HILSE, Hans-Werner, KOTHE, Jochen. *Implementing Persistent Identifiers*. Consortium of European Research Libraries, 2006. Disponible sur: http://webdoc.sub.gwdg.de/edoc/ah/2006/hilse_kothe/urn%3Anbn%3Ade%3Agbv%3A7-isbn-90-6984-508-3-8.pdf [consulté le 12/04/2017]

[405] ARCHER, Phil, GOEDERTIER, Stijn, LOUTAS, Nikolaos. *Op. Cit.*

[406] BERNERS-LEE, Tim. *Uniform Resource Identifier (URI): Generic Syntax. Op. Cit.*

[407] *Ibid*.

En outre, le choix d'une URI opaque ou signifiante est important :

- Les identifiants opaques sont intéressants car la sémantique a tendance à « pourrir »[408] avec le temps, et à être difficilement gérable lorsqu'il y a plusieurs langues. De plus, ces identifiants peuvent être créés facilement de manière automatique et sont plus opérationnels lorsque l'on traite de grands jeux de données qui ont de hautes probabilités d'occasionner du « conflit » et de l'ambiguïté.
- Les identifiants signifiants sont manipulables, ils ont des métadonnées *ad hoc*, visibles et exploitables immédiatement. Ils sont plus indiqués pour les petits jeux de données ou des URL en contact avec l'utilisateur directement (nous l'avons vu avec la direction des Archives de France, par exemple).

> « Il faut toujours se souvenir que les identifiants sont censés être traités à la fois par des machines et à la fois par des humains : ils devraient pouvoir être lus, épelés et écrits manuellement, mais également traités par des ordinateurs. »[409]

Certains identifiants tels que les identifiants ARK tentent de préserver les bénéfices de la signifiance et de l'opacité tout en retirant les contraintes [410]: la proximité de l'identifiant opaque ARK avec ses métadonnées (il suffit par exemple de rajouter un « ? » à la fin lors du déréférencement pour obtenir celles-ci) permet d'obtenir une certaine souplesse. Les identifiants identifient mais ne définissent pas : ils dénotent, ils réfèrent, ils pointent vers quelque chose. Ce sont les métadonnées qui doivent jouer le rôle de descripteur.[411]

Ils ne sont cependant pas des dénominations non plus. « *Associer absolument des URI à une fonction de nom est trop lourd en plus d'être incorrect la plupart du temps* ».[412] L'idée perdure que le nom de domaine (et par extension les identifiants web) ont une valeur de propriété et sont « équivalents à une valeur immobilière ». Les rachats de nom de domaine se comptent en millions (cas de pizza.com, par exemple, dont le prix est évalué à 2,6millions de dollars…[413]) et si le *cybersquatting** sur des marques déposées est illégal, il y a quand même une économie qui se maintient autour du rachat et de la vente des noms. Ainsi, individualiser l'identifiant et le séparer du nom de domaine n'est pas une mauvaise idée, cela règle le problème évoqué par Ian Davis dans son article *Is 303 really necessary ?* : si un serveur web cesse d'exister certaines informations seront perdues.

Afin de condenser ces quelques prérequis, ARK propose dans sa structure un lien entre trois composants : l'objet, ses métadonnées et une déclaration d'engagement. Ici la vision de l'identifiant est donc plurielle et tente d'englober l'ensemble des caractéristiques évoquées dans notre première partie : dénomination, localisation, contextualisation. C'est ce que l'on souhaite faire avec nos URI.

---

[408] *To rot* en anglais, devenir obsolète, qui est lié au concept de *link rot* (phénomène d'obsolescence progressive des liens qui deviennent des liens morts relativement utilisé dans la littérature professionnelle).

[409] HILSE, Hans-Werner, KOTHE, Jochen. *Op. Cit.*

[410] KUNZE, John A. *Op. Cit.*

[411] BERGMAN, Mike. *Op. Cit.*

[412] *Ibid.*

[413] REES, Mark. *Nom de domaine : 2,6 millions de dollars pour pizza.com*. Site Next inpact.com [en ligne] Disponible sur https://www.nextinpact.com/archive/42850-pizzacom-sedo-nom-domaine-record.htm [consulté le 09/08/2017]

3.3.3. Conseil n°3 : Assigner, « mapper » et penser le déréférencement

Une fois les identifiants définis, il s'agit de procéder à l'attribution même des identifiants, c'est-à-dire créer concrètement le lien entre l'URI et la ressource (John Kunze dirait : créer concrètement l'identifiant car, nous l'avons vu, pour lui l'identifiant doit être la symbolisation même de ce lien). L'assignation doit être faite le plus tôt possible dans la vie d'un contenu. Deux types de stratégie peuvent alors être mis en place[414] :

- La stratégie *big bang* : générer en une seule fois tous les identifiants,
- La stratégie progressive : avancer par sous-ensembles homogènes (application métier par application métier, type de ressource par type de ressource…).

En effet, une fois la structure de l'identifiant construite (à l'étape précédente), il est possible de créer les identifiants à la volée à partir de celle-ci, manuellement ou automatiquement. L'auto-incrémentation des entrées lors de l'attribution peut être intéressante, mais il n'est conseillé par les auteurs de l'ISA que dans deux cas : celui dans lequel le procédé ne sera jamais répété, et celui dans lequel le procédé sera répété pour recréer exactement les mêmes identifiants pour les mêmes données (et à condition que les nouvelles URI ne soient attribuées qu'aux nouveaux éléments de données ajoutées)[415]. Dans tous les cas il fortement conseillé de tester préalablement l'attribution sur un échantillon représentatif, permettant des vérifications techniques.

Utiliser un répertoire d'attribution peut permettre de fixer des cadres et procédures. La gouvernance de ce répertoire peut être soit établie grâce à des standards, soit être propriétaire, et elle doit permettre de mettre en place la confiance envers la pérennité et la véracité de ses identifiants ainsi que de ses métadonnées.[416]

Pour Tim Berners-Lee, les prérequis d'une URL sont que toutes les représentations doivent contenir les mêmes informations. Ce contenu, cette forme, ce travail conceptuel intègre est ce qui est identifié par l'URI en question. Mais, nous l'avons vu, il peut y avoir différents niveaux de granularité dans l'attribution d'identifiants : un identifiant « mère » qui va désigner le concept tel qu'imaginé par Tim Berners-Lee, et des identifiants spécifiques plus détaillés, pour les différentes représentations par exemple. Cela est nécessaire notamment dans les cas où les structures n'implémentent pas de négociation de contenu.

La gestion du co-référencement (plusieurs identifiants pour un même référent) et la gestion de l'ambigüité (plusieurs référents pour un identifiant) ne se font pas du tout de la même façon : si le co-référencement peut éventuellement être un outil souhaitable (nous l'avons vu avec les Archives de France), l'ambiguïté est à proscrire impérativement. De même, la question de la réassignation d'identifiant peut se poser à l'aulne des avancées actuelles : en tout cas en 2014 Norman Paskin et Godfrey Rust ne recommandaient absolument pas la pratique.[417] Le lien entre les ressources doit donc être réfléchi : les différentes représentations d'un élément doivent être « mappées », réutiliser quasiment le même nom ne suffit pas. Cela se fait au moyen du modèle de données mis en place, en précisant les types de relation

[414] PEYRARD, Sébastien, TRAMONI, Jean-Philippe, KUNZE, John. *Op. Cit.*

[415] ARCHER, Phil, GOEDERTIER, Stijn, LOUTAS, Nikolaos. *Op. Cit.*

[416] PASKIN, Norman, RUST, Godfrey. *Op. Cit.*

[417] *Ibid.*

constitutives du lien entre deux entités : partie-tout, co-référence, alternative, etc. Ce type de déclaration d'équivalence se trouve dans la plupart des ontologies. [418]

En ce qui concerne la résolution, Sébastien Peyrard, Jean-Philippe Tramoni et John Kunze apportent trois éléments de réponse quant à la manière de mener le déréférencement des ressources [419]:

- Le déréférencement par réécriture, qui répond au problème de l'identification des pages web dynamiques qui « assemblent » des contenus sur le vif pour ressembler aux pages web statiques. Cela peut se faire en demandant au serveur la réécriture des URL à travers des règles précises, telles que l'interdiction de réutiliser le titre de la page systématiquement, ou choisir des éléments qui sont intéressants pour la pérennité de l'URL, par exemple. C'est également le système utilisé pour les redirections 303 dans le cas du déréférencement d'URI d'objets réels.
- Le déréférencement par réédition, en déterminant une URL dédiée pour la citation de la ressource, un permalien. Par exemple, il est possible d'avoir une URL fixe pour un contenu mais qui redirige toujours vers la version la plus récente du document. Ainsi, on n'utilise pas le nom de domaine de la localisation réelle de la ressource, mais un nom de domaine tiers.
- Le déréférencement par résolveur interne. Il s'agit de faire correspondre une URL à un objet, via un logiciel de résolution d'identifiant (par exemple Handle). Cela nécessite un gros travail d'implémentation mais permet une souplesse, notamment pour les mises à jour ou en cas d'évolution des ressources.

Ainsi, on distingue deux types d'URI : les URI abstraites qui identifient la dernière version d'une ressource, et les URI concrètes qui identifient une version particulière de la ressource identifiée.[420] Lors de l'assignation, la gestion de ces deux types d'identifiants est primordiale, afin d'adapter l'implémentation et construire une gestion logique. Comme nous l'avons vu, la résolution en http (requête http GET) est nettement privilégiée par la communauté professionnelle. Quant au choix de la gestion des identifiants d'objets réels, la redirection 303 offre plus de flexibilité mais les URI Hash peuvent se révéler utiles lorsque l'on a un accès limité à la configuration du serveur. Le choix de l'un ou de l'autre n'impacte pas de toute manière les utilisateurs[421].

3.3.4. Conseil n°4 : Enrichir, déployer, gérer et maintenir

Nous n'avons pas encore évoqué un élément crucial du développement d'un projet d'identification : les métadonnées, pierres angulaires de ces systèmes, qui permettent le maintien des liens.

La gouvernance s'applique aussi aux métadonnées, et celles-ci doivent être en nombre suffisant afin d'assumer leur rôle de désambiguïsation. Les métadonnées principales doivent de préférence être publiées en formats extensibles et interopérables (XML, RDF-TTL, JSON) et utiliser les modèles de données et vocabulaires contrôlés lorsque cela est approprié. Assignées par une autorité

[418] Nous n'irons pas en profondeur sur l'implémentation d'un modèle de données car cela sort de notre propos.

[419] PEYRARD, Sébastien, TRAMONI, Jean-Philippe, KUNZE, John. *Op. Cit.*

[420] *Ibid.*

[421] DAVIDSON, Paul, CIO Sedgemoor, District Council. *Op. Cit.*

compétente, elles doivent également apporter la garantie qu'elles n'ont pas été altérées depuis leur création.[422]

Un autre élément important est l'indication de leur provenance : rappelons-nous les catégories de producteurs de métadonnées définies par Mirna Willer et Gordon Dunsire, sont-elles donc définies par un professionnel de l'information, nous garantissant ainsi une métadonnée du meilleur cru ? Proviennent-elles d'un amateur de l'information, qui pourrait tout aussi bien avoir menti, s'être montré négligent, avoir fait des erreurs lors de leur attribution [423]? Ou, pis encore, sont-elles le fruit d'un processus automatique ? Cette information est essentielle et conditionne aussi la possibilité de réutiliser les données que l'on publie.

David Wood propose d'ailleurs des solutions point par point très intéressantes pour contrer l'argumentaire de Corey Doctorow[424], et ce afin d'optimiser la création de métadonnées participatives[425] :

- Les gens mentent : dans ce cas permettons-leur de se baser sur un modèle de confiance,
- Ils sont feignants : automatisons quand c'est possible et encourageons « l'*authoring** »,
- Ils sont ignorants : automatisons la vérification lorsque c'est possible,
- Il n'est pas possible de se connaitre soi-même : permettons qu'il y ait plusieurs sources de métadonnées,
- Les schémas ne sont jamais neutres : permettons qu'il y ait plusieurs schémas possibles,
- Les statistiques influencent les résultats : permettons qu'il y ait plusieurs points de vue coexistant,
- Il y a plus d'une façon de décrire quelque chose : permettons qu'il y ait plusieurs descriptions.

Ces idées démontrent tout à fait que le déploiement d'un système d'identification nécessite une gestion vraiment souple et modulable face au changement. En effet, la flexibilité est aussi un critère qui participe à assurer l'interaction entre les systèmes.

Quant à la maintenance, faire vivre un système d'identification est complexe, et demande un investissement au moins aussi important que celui qui a permis sa mise en place. La technologie n'apporte pas la pérennité, mais la pérennité est néanmoins assistée par la technologie. « *Tout changement de lieu, de droits ou tout autre métadonnée doit être répercuté dans le système d'identification. En conséquence, chaque migration de la base documentaire (par exemple vers un nouveau serveur) implique du travail de maintenance au sein du système.* »[426]. Il s'agit, surtout, de marquer les changements et les évolutions :

- Indiquer le *versioning* (par exemple, une route que l'on va faire agrandir, une ressource qui va être actualisée),

[422] PASKIN, Norman, RUST, Godfrey. *Op. Cit.*

[423] DOCTOROW, Cory. *Op. Cit.*

[424] *Ibid.*

[425] WOOD, David. *Op. Cit.*

[426] HILSE, Hans-Werner, KOTHE, Jochen. *Op. Cit.*

- Préciser le statut (par exemple, un document non signé, un document validé),
- Proposer des URI qui mènent à des versions plus récentes, etc.

Créer des URI de représentation permet notamment l'accès direct aux ressources sans changer les informations d'un *header* dans une requête, ce qui est particulièrement utile. En outre, chaque représentation de ce même document devra pointer vers les autres pour faciliter le cheminement. Il faut tout de même toujours avoir en tête la notion suivante : qui dit maintenance, dit future migration. Si l'identifiant correctement construit ne perd jamais sa fonction d'identification, la structure sous-jacente, elle, peut cesser d'exister. Il y aura forcément ce type de traitement à effectuer tôt au tard, et posséder un système efficace, simple et qui se tient peut faciliter grandement ces opérations.

Pour conclure sur l'utilisation des données liées, Chris Bizer et Tom Heath souhaitent tout de même préciser que celles-ci ne remplacent pas une infrastructure de management de données, elles la complètent. La clé est de se baser sur les standards en vigueur et réutiliser ce qui a déjà été fait[427]. En 2004 déjà, les résultats du séminaire de l'ERPA à Cork sur les systèmes d'identifiants pérennes préconisaient une meilleure communication entre les gestionnaires de projet et la recherche et développement car cela devait permettre l'interopérabilité des schémas de nommage, à la fois entre projets, pays, mais également entre disciplines. L'objectif global à atteindre nécessite en effet une implication de l'ensemble des acteurs, parce qu'elle permet de profiter du meilleur de chaque expérience et de satisfaire l'objectif final, la satisfaction des besoins utilisateurs.

[427] BIZER, Christian, HEATH, Tom. *Op. Cit.*

Conclusion de la partie 3

La partie 3 nous a enseigné que :

- Les structures définissent à travers leurs projets d'identification des objectifs très liés à leurs aspirations profondes, et pas uniquement à leurs missions premières comme nous pourrions le penser à priori,
- Une stratégie donnée peut être une très mauvaise idée pour une structure et constituer la solution idéale pour une autre,
- Le principal n'est pas d'obtenir un identifiant aux caractéristiques idéales, mais bien de s'assurer que les buts fixés sont atteints, quels que soient les moyens employés et la manière utilisée,
- La pérennité n'est pas une question de technique, elle est un engagement « administratif » qui implique tous les acteurs de la chaîne, mais qui s'appuie tout de même sur une implémentation technique,
- Il ne faut pas dénigrer les méthodes de création de métadonnées folksonomiques ni les « foyers » de connaissances participatives, ils ont un bel avenir devant eux et peuvent être d'une grande aide dans les projets pour peu qu'on les utilise de manière adéquate,
- Assurer en amont l'interopérabilité facilite grandement l'atteinte des objectifs finaux,
- Les choix faits en début de projet, notamment la construction de la syntaxe de l'identifiant et ses caractéristiques (opacité, signifiance) conditionnent les possibilités de manipulation future,
- La gouvernance s'applique à tous les aspects du projet : structuration de l'identifiant, création de métadonnée, gestion, maintenance… Elle est primordiale et doit être pensée en amont,
- La communication entre les acteurs, interne ou externes est très importante et les données liées ne sont efficaces et utiles que si les acteurs (humains) sont déjà « liés ».

Pour conclure, lorsque l'on identifie, il ne faut surtout pas perdre de vue les objectifs initiaux qui ont motivé la mise en place du projet. En effet, à vouloir déployer un système idéal qui conviendra à tout le monde il est possible de ne plus distinguer le but ultime du moyen d'accéder à ce but. Globalement, les buts premiers sont l'expérience utilisateur, l'accessibilité, l'efficacité, la pérennité et la praticité. Quant à l'interopérabilité, l'universalité, la granularité, l'unicité, l'échelle, l'adaptabilité, ce sont des moyens de parvenir à ces buts, ils ne devraient pas constituer une fin en soi. La mise en place de tels projets nécessite donc une définition stricte du périmètre et une vision sur le long terme.

CONCLUSION

Comment utiliser l'identification pour optimiser les bénéfices des données liées et s'adapter aux enjeux des paradigmes actuels ? C'est la question qui a été présente en filigrane tout au long de l'écriture de ce mémoire.

La problématique pourrait bien sûr être résolue avec des conseils pratiques, des dispositions techniques et des recommandations pointues formulées par des spécialistes de la question. Toutefois, si nous avons essayé de citer tout au long de notre développement quelques points de vue apportés par des personnes et des organisations référentes en la matière, il est cependant hors de notre propos et de notre portée d'y opposer un point de vue subjectif ou des critiques. Des éléments de cette problématique, parmi lesquels ceux évoqués en introduction, peuvent néanmoins être discutés depuis un poste d'observation théorique. Cette conclusion sera donc développée à travers l'exploration de trois constatations qui se sont imposées au cours de cette recherche.

De l'adaptabilité du vivant, appliquée aux systèmes

La première constatation touche à notre problématique globale : comment créer un système d'identification optimal ? Il semble que la réussite d'un système d'identification passe par le respect de quatre critères très importants, inter-liés entre eux. Parmi ceux-ci, un critère primordial et un seul, occulte tous les autres : la capacité à s'adapter. Cette qualité, qui est d'ailleurs la condition même de la survie du vivant, s'applique en effet relativement bien de manière générale à la création humaine qu'est la technologie. Les cas que nous avons étudiés ont un point commun : ils nous apprennent tous d'une façon ou d'une autre que la réussite d'un système (*in fine* qu'il soit technique ou biologique) tient à la capacité qu'il a à s'adapter à son environnement, ou à le modifier. Après tout, la nature elle-même nous enseigne cette leçon depuis des millénaires d'évolution, ce qui s'adapte survit, ce qui s'obstine meurt[428].

S'adapter et interagir avec un environnement, c'est avant tout créer une communication avec lui, bi-directionnellement : le comprendre et l'influencer par ses propres changements ; ou uni-directionnellement : simplement l'appréhender et s'y conformer pour pouvoir se développer en son sein. La notion de communication, employée dans le cas de systèmes qu'ils soient vivants ou technologiques, est particulièrement parlante, tout simplement parce qu'elle est une condition de base au développement de l'adaptabilité. Dans le domaine spécifique de la technologie et de l'ingénierie, l'Association Française d'Ingénierie Système (AFIS) définit ainsi la notion de système :

> « Un système est décrit comme un ensemble d'éléments en interaction entre eux et avec l'environnement, intégré pour rendre à son environnement les services correspondant à sa finalité. »[429]

[428] Et pour citer La Fontaine, le conte du chêne et du roseau est une bonne métaphore explicitant cette idée.

[429] AFIS. *Le système et sa définition.* Afis.fr [en ligne] Disponible sur https://www.afis.fr/nm-is/Pages/Ing%C3%A9nierie%20Syst%C3%A8me/Le%20syst%C3%A8me%20et%20sa%20d%C3%A9finition.aspx [consulté le 16/08/2017]

Remarquons dans cette citation l'importance de ces deux notions de communication et d'environnement, qui montrent bien comment le besoin d'interagir est partagé, à la fois dans les fonctions primaires du vivant (nutrition, *relation*, reproduction) et dans les systèmes technologiques (se placer dans un environnement, communiquer en interne et en externe). Ce sont deux mécaniques différentes qui fonctionnent pourtant sur un principe commun : la capacité à communiquer et s'enrichir mutuellement avec ce qui l'entourent. Joël Bockaert, neurobiologiste et membre de l'Académie des Sciences, développe bien cette idée dans son ouvrage au titre évocateur *La communication du vivant, de la bactérie à Internet*. Il déclare que « *la communication est de fait si essentielle à la vie et à son évolution, de la bactérie aux sociétés humaines, que l'on peut se demander s'il ne faut pas substituer au « je pense donc je suis » de Descartes un « je communique donc je suis » […]* ».[430]

Les identifiants sont des produits de cette logique à partir du moment où ils sont présents dans les aspects structurels d'un système. Ils doivent être pensés pour atteindre cet objectif d'adaptabilité, et ils en constituent à la fois le moyen et le but final.

L'adaptabilité engendre donc des notions connexes sans lesquelles elle ne peut être garantie. Nous les avons vues : la communication interne et externe, l'interopérabilité, la souplesse et la pérennité (qu'elle soit une pérennité d'accès, de structure ou tout simplement une notion de persistance dans le temps). Cette adaptabilité doit s'étendre sur plusieurs dimensions :

- Bien entendu, l'adaptabilité à l'environnement et à son contexte ;
- La capacité à se comprendre soi-même et harmoniser son propre fonctionnement interne ;
- La capacité à interopérer avec les autres, se nourrir de leurs forces et apprendre de leurs faiblesses (en acceptant parfois de fusionner et construire ainsi une structure plus importante[431]) ;
- Et enfin, surtout, la capacité à être rétroactif, c'est-à-dire savoir s'adapter à son propre parcours et rester en cohérence avec lui.

Après tout, « *les espèces qui survivent ne sont pas les espèces les plus fortes ni les plus intelligentes, mais celles qui s'adaptent le mieux aux changements.* »[432]

Le web de données et son impact actuel : neutralité, économie et droit de propriété

La seconde constatation découle de la question de l'intérêt de la mise en place des données liées pour une structure, et par extension de l'intérêt d'être intégré et d'intégrer ses ressources dans le web de données.

---

[430] BOCKAERT, Joël. *La communication du vivant : de la bactérie à Internet*. O. Jackob, 2017. 205p.

[431] Pour étendre la métaphore du vivant, faisons un rapprochement avec la *phagocytose*: capturer et ingérer un organisme.

[432] Phrase gravée dans le marbre à l'Académie des Sciences de Californie en citation à Darwin, qui ne serait d'ailleurs pas de Darwin du tout. HOOD, Marlowe. *Bicentenaire de Darwin : les citations qu'il n'a jamais dites*. Lapresse.ca, 2009. [en ligne] Disponible sur http://www.lapresse.ca/sciences/200902/12/01-826782-bicentenaire-de-darwin-les-citations-quil-na-jamais-dites.php [consulté le 31/08/2017]

La feuille de route stratégique du ministère de la culture et de la communication sur les *métadonnées culturelles et la transition web 3.0* comporte commodément une liaison avec notre métaphore biologique :

> « Le rebond de données en données organiquement liées entre elles constitue un véritable écosystème culturel vivant. »[433]

Economiquement, il est intéressant de remarquer que le web de données (et par extension, l'identification permettant la création d'un graphe global géant) apporte un avantage important aux institutions culturelles sur un domaine, le web, qui pour l'instant leur échappait quasi-totalement : elles avaient énormément de mal à s'intégrer dans le numérique où les acteurs « économiques » trouvaient un terrain d'activité florissant. Prenons le cas assez parlant des bibliothèques : en effet, la profusion d'informations sur le web donnant accès à « tout le savoir du monde » laissait à penser que les bibliothèques étaient révolues, dépossédées de leur raison d'être…[434] Des nouveaux acteurs économiques se pressaient pour développer des systèmes imitant les bibliothèques, en constituant des bases de connaissances et de biens informationnels mises en ligne sur le web. Le monde de la propriété évolue en un monde de l'accès[435].

Certains de ces nouveaux acteurs restent sur le principe de la gratuité, et trouvent des moyens de rémunération adjacents. Par exemple, Spotify et Deezer pour la musique proposent des services typiques des bibliothèques : un accès illimité et gratuit, qu'ils vont compléter par un ajout de services payants. Leur popularité et l'utilisation de plages de publicité sur les comptes non abonnés permet de rentabiliser un système qui s'approprie le principe de la bibliothèque en ce qui concerne leur politique d'accès.

Certains acteurs proposent quant à eux la monétisation de l'accès même : fonctionnant avec ce que l'on appelle des *paywall*s, bloquant l'accès à l'aide d'un système de paiement. Il s'agit d'un système qui tend à être très restrictif, très controversé, et qui engendre des mouvements « de rébellion » face à l'hégémonie des fournisseurs de contenus[436].

Dans le cas d'autres acteurs culturels comme la BBC, d'autres types de « concurrents » tels que les fournisseurs de vidéo à la demande, les sites de *replay* des chaines privées, etc. peuvent également être pris en compte. Pour les archives, même s'il n'y a clairement pas de menace concurrentielle directe, il y a un intérêt fort à rendre plus accessibles en ligne au grand public des contenus pour valoriser les fonds. Le système économique ne peut décidément pas se baser sur une répression permanente : les bouleversements en faveur de *l'open-access* (notamment

---

[433] Ministère de la Culture et de la Communication. *Métadonnées culturelles et transition Web 3.0, Feuille de route stratégique*. 2014. [en ligne] Disponible sur https://www.inha.fr/_attachments/le-web-semantique-pour-les-donnees-culturelles-actualite/64776-feuille-de-route-strategique-metadonnees-culturelles-et-transition-web-3-0%25282%2529.pdf?download=true [consulté le 22/03/2017]

[434] BERTRAND, Anne-Marie. *Les bibliothèques*. La Découverte, 2011. 128p. p111.

[435] RIFKIN, Jérémy. *L'âge de l'accès, la nouvelle culture du capitalisme*. La Découverte, 2005. 406p.

[436] Nous citerons notamment le cas de Sci-Hub et LibGen, plateformes pirates divulguant des publications scientifiques accessibles normalement uniquement via les fameux « *paywalls* » des éditeurs scientifiques. Ces derniers mènent une guerre sans merci contre ce type d'initiative *open-access*, à en juger par le tout dernier délibéré de la cour de justice de New-York les condamnant chacun à verser 15 millions de dollars à Elsevier. Source : Sciences et avenir.fr [en ligne] https://www.sciencesetavenir.fr/fondamental/le-proces-des-pirates-condamne-sci-hub-15-millions-de-dollars_114269 [consulté le 16/08/2017]

pour les articles scientifiques) privent déjà les éditeurs de leur marché et ceux-ci seront donc tôt ou tard obligés de se repositionner.

Avec *l'open-data*, *l'open-access* et les données liées, les institutions culturelles trouvent donc dans le numérique un nouveau moyen de perdurer encore et toujours dans leurs missions premières et de retrouver une position de fournisseurs de contenus prisés. Nous l'avons également vu avec le LED (*Linked Enterprise Data*), le principe des données liées apporte aussi des bénéfices et des fonctionnalités intéressantes pour les entreprises en dehors du web, elles participent à créer un système d'information hyper-connecté, « intelligent », qui est capable de fournir du contexte et des recommandations spécifiques. Cela mobilise davantage d'acteurs très impliqués, notamment grâce aux possibilités qu'offre l'indexation folksonomique :

> «La montée en puissance du web 2.0 et du web sémantique renforce des possibilités d'échanges de données jusqu'à lors insoupçonnées en permettant une classification et une structuration des contenus informationnels par la mobilisation de l'intelligence collective, c'est-à-dire de la puissance créatrice et inventive des individus. »[437]

De plus, le partage des données a aussi un impact sur les principes fondamentaux des institutions culturelles. C'est le cas de la neutralité par exemple, en bibliothèques comme en archives, ou dans toute structure culturelle, qui est un fondement très discuté auxquels de nombreux chercheurs ont tenté d'apporter des réponses[438]. Or la neutralité est très liée à la notion d'exhaustivité : comment être neutre si on ne présente qu'une partie des informations ? Ne présenter qu'une partie des informations, c'est faire une sélection, faire une sélection c'est faire des choix, et faire des choix, c'est être partial. Le web de données, qui promet une globalisation de l'information humaine disponible et accessible sans conditions, participe à l'atteinte de ces objectifs de neutralité et d'impartialité. Les institutions culturelles, en s'impliquant dans ce type de projet, marquent fortement leur volonté de s'inscrire dans une appréhension neutre des savoirs, en proposant une réutilisation participative de ces données.

Par extension, l'identification est donc au cœur de ces nouveaux enjeux qui conditionnent l'utilisation des ressources culturelles sur le web, notamment sur le web de données.

Du numérique et de ses apports au monde de l'identification

La troisième et dernière constatation ayant émergé lors de cette recherche concerne notre questionnement sur le numérique et les enjeux et conséquences économiques, techniques et culturelles. Replacer dans son contexte l'identification à l'heure où le numérique n'existait pas a en effet permis de révéler des postulats infondés.

Nous avions comme hypothèse de départ l'idée que l'identifiant physique et ses principes, réutilisés, pouvaient permettre d'apporter du bon sens et de l'organisation à l'identification numérique, voire de résoudre ses problèmes. En effet, il semblait possible d'imaginer que, de manière générale, la transposition d'un

[437] Ministère de la Culture et de la Communication. *Métadonnées culturelles et transition Web 3.0, Feuille de route stratégique*. *Op. Cit.*

[438] Nous citerons notamment les travaux de Jean-Luc Gautier-Gentès, ou encore Raphaëlle Bats.

système longuement travaillé au sein d'un environnement ancien et stable à un nouvel environnement devrait pouvoir permettre de ne pas tout recommencer : l'on aurait pu ainsi postuler que le numérique bénéficie des apprentissages passés en termes d'identification physique. Globalement il semblerait que la supposition ce soit totalement retournée sur elle-même : des problèmes sont nés de cet entêtement qu'il y a eu à vouloir appliquer au numérique des principes papier (la crise identitaire, le *http range 14*, le *link maintenance problem*, l'opacité des URL... sont des exemples intéressants montrant comment notre logique de bureau, fichiers, dossiers, est source de confusions), et c'est l'apport du numérique et des nouveaux points de vue qui permet de trouver un éclairage sur nos pratiques ancestrales en matière de cotation. Pour le dire simplement, l'un va profiter à l'autre mais pas dans le sens où l'on pourrait le croire de prime abord.

Ainsi, le numérique apporte réellement des réponses qui vont permettre de faire avancer des problématiques en débat depuis des siècles dans les communautés professionnelles en termes de cote. Pour prendre l'exemple des archives que nous avons développé précédemment, les systèmes de cotation archivistiques papier pourraient vraiment profiter des lumières de l'identification de données uniques et pérennes sur le web. En effet, de nouvelles conceptions de la cotation, notamment la cotation multiple proposée par Baptiste De Coulon, émergent de ces constatations faites à travers les projets de mise en ligne sur le web de données. Dans le cas des bibliothèques, les cotations bibliographiques profitent déjà de l'avancée de la globalisation des données et de *l'open-access*, avec des liens systématiques effectués entre les identifiants de localisation et ceux de singularisation globalisés (la notice est l'outil qui permet d'en rassembler les différentes coréférences au sein des systèmes).

Ainsi, le changement de paradigme permet de développer des aspects inédits et surtout d'avoir un regard neuf sur la définition de ce qu'est un document, une information, un identifiant.

Pour clore notre propos, nous préciserons que le paradigme numérique évolue chaque jour. Chaque institution peut apprendre de ses erreurs et évoluer pour devenir un jour une référence dans son domaine. Les choses ne sont jamais figées : il reste des systèmes à découvrir et de nouvelles expériences à retirer des futurs projets. L'identification est un sujet complexe et passionnant qui mérite d'être placé au centre des débats, car il cristallise beaucoup de problématiques actuelles liées au web. C'est, en effet, un sujet foisonnant qui a encore un bel avenir devant lui.

BIBLIOGRAPHIE

Identification des objets physiques

Bibliothèques

BERTRAND, Anne-Marie. *Les bibliothèques*. La Découverte, 2011. 128p. p111.

MENON, Bruno. *Web et bibliothèques, entre métaphore et mimésis*. Journée d'étude du groupe TICIS-SFSIC : Le web a-t-il un sens ? Paris, 2010. [en ligne] Disponible sur https://hal.archives-ouvertes.fr/halshs-00647868/document [consulté le 16/08/2017]

OTLET, Paul. *Traité de documentation*. Liège : C.L.P.C.F, 1989. Réimpression de l'édition de 1934. Préface de Robert ESTIVALS, Avant-propos d'André CANONNE.

SAKALAKI, Maria, THEPAUT, Yves. La valeur de l'information, évaluation des biens informationnels versus bien matériels. *Questions de communication*, 2005. N°8. [en ligne] Disponible sur https://questionsdecommunication.revues.org/5300#quotation [consulté le 14/07/2017]

VIRY, Claude-Michel. *Guide historique des classifications de savoirs ; enseignement, encyclopédies, bibliothèques*. L'Harmattan, Paris, 2013. 256 p.

Archives

AAF. *Abrégé d'archivistique, Principes et pratiques du métier d'archiviste*. 2012.

DE COULON, Baptiste. *De l'intérêt de la cotation multiple en archivistique*. Site siar.hypotheses.org, 2016. [en ligne] Disponible sur https://siar.hypotheses.org/59 [consulté le 09/05/2017]

Direction des archives de France, *La Pratique archivistique française*, Paris, 2008.

GARON, Louis. A-t-on besoin d'une nouvelle définition de l'article ? *Archives*, n°21, 1989. p83-87.

HEON, Gilles. L'article dans les répertoires : élément de cotation ou élément de rangement ? *La Gazette des archives*, n°136, 1987. p5-16. [en ligne] Disponible sur http://www.persee.fr/doc/gazar_0016-5522_1987_num_136_1_3018 [consulté le 06/07/2017]

Musées

GOB, André, DROUGUET, Noémie. *La Muséologie*, Armand Colin, 4e édition, 2014.

MUSEE DES BEAUX-ARTS DE LYON. *Le récolement*. Article sur le site du Musée des Beaux-Arts de Lyon, date inconnue [en ligne]. Disponible sur http://www.mba-lyon.fr/mba/sections/fr/collections-musee/vie-des-collections/le-recolement [consulté le 01/06/2017]

ICOM, *Norme ObjectID*, Site Archives ICOM Muséum.fr [en ligne] disponible sur : http://archives.icom.museum/objectid/how_fr.html [consulté le 27/04/2017]

Edition et publications

HONORE, Suzanne. *La numérotation normalisée internationale du livre (International Standard Book Number.)* Site du Bulletin des Bibliothèques de France [en ligne] http://bbf.enssib.fr/consulter/bbf-1969-08-0321-001 [consulté le 27/04/2017]

RIFKIN, Jérémy. *L'âge de l'accès, la nouvelle culture du capitalisme*. La Découverte, 2005. 406p.

Le Web sémantique et le web de données

ALIX, Yves (dir.). *Bibliothèques en France 1998-2013*. WENZ, Romain. Chapitre « Des catalogues aux métadonnées : la bibliothèque vers le Web sémantique » p.160-171. Editions du cercle de la librairie, 279p. 2013.

BERGMAN, Mike. *Give me a sign: what do things mean on the semantic web?* Mkbergman.com, 2012. [en ligne] Disponible sur: http://www.mkbergman.com/994/give-me-a-sign-what-do-things-mean-on-the-semantic-web/ [consulté le 18/05/2017]

BERMÈS, Emmanuelle, ISAAC, Antoine, POUPEAU, Gautier. *Le Web sémantique en bibliothèque*. Paris : Electre-Ed. du Cercle de la Librairie, 2013.

BERNES-LEE, Tim, HANDLER, James, LASSILA, Ora. The semantic web. *Scientific American Magazine*. 2001.

BIZER, Christian, HEATH, Tom. *Le web de données*. Pearson France, 2012. [en ligne] Disponible sur https://www.pearson.fr/resources/titles/27440100179400/extras/4746_chap03.pdf [consulté le 09/05/2017]

BRATT, Steve. *Semantic Web, and other technologies to watch.* W3C, 2007. [en ligne] Disponible sur https://www.w3.org/2007/Talks/0130-sb-W3CTechSemWeb/0130-sb-W3CTechSemWeb.pdf [consulté le 29/11/2016]

DOCTOROW, Cory. *Metacrap: Putting the torch to seven straw-men of the meta-utopia*. 2001. [en ligne] Disponible sur https://www.well.com/~doctorow/metacrap.htm [consulté le 06/07/2017]

HYVONEN, Eero. *Publishing and Using Cultural Heritage Linked Data on the Semantic Web*. 2012

ISAAC, Antoine. *Les référentiels: typologie et interopérabilité*. Séminaire IST Inria : le document numérique à l'heure du web de données, Carnac 2012. [en ligne] Disponible sur https://hal.inria.fr/hal-00740282v1 [consulté le 09/05/2017]

MILNE, David, MEDELYAN, Olena, WITTEN, Ian H. *Mining Domain-Specific Thesauri from Wikipédia : a case study*. Department of Computer Science, University of Waikato. Hamilton, 2006. [en ligne] Disponible sur http://www.cs.waikato.ac.nz/~ihw/papers/06-DM-OM-IHW-wikipedia_vs_agrovoc.pdf [consulté le 25/08/2017]

OURY, Clément. *ISSN : Transitioning to linked data. Data in library: the big picture*. Satellite Meeting of IFLA World Library and Information Congress, Chicago, 2016. [en ligne] https://halshs.archives-ouvertes.fr/halshs-01358415/document [consulté le 22/03/2017]

PAN, Jeff Z., HORROCKS, Ian. *Metamodeling Architecture of Web Ontology Languages*. University of Manchester, 2001. [en ligne] Disponible sur http://www.cs.man.ac.uk/~horrocks/Publications/download/2001/rdfsfa.pdf [consulté le 31/08/2017]

PEYRARD, Sébastien, SIMON, Agnès. Le web sémantique en bibliothèque. *Bulletin des Bibliothèques de France*, 2014. p189-191. [en ligne] Disponible sur http://bbf.enssib.fr/consulter/bbf-2014-02-0189-007

VAN HOOLAND, Seth. VERBORGH, Ruben. *Linked Data for Libraries, Archives and Museums: How to Clean, Link and Publish your Metadata*. Amer Library Assn Editions, 2014. 224p.

Web sémantique et web de données, Sensibilisation à l'évolution des catalogues. Programme Transition bibliographique, Réseau national des formateurs, 2016. [en ligne] Disponible sur https://www.transition-bibliographique.fr/wp-content/uploads/2016/04/Web_de_Donnees_26-02-2016_Version_Courte.pdf [consulté le 16/07/2017]

WENZ, Romain. Hypertextualisation, La quête du lien sémantique en bibliothèque. *Revue de la BNF*. Bibliothèque nationale de France, 2012. N°42. p36-41

WILLER, Mirna, DUNSIRE, Gordon. *Bibliographic Information Organization in the Semantic Web, 1st Edition*. Chandos Publishing, 2013. 350p.

W3C, Library Linked Data Incubator Group Final Report. W3C Incubator Group Report, 2011. [en ligne] Disponible sur https://www.w3.org/2005/Incubator/lld/XGR-lld-20111025/ [consulté le 18/02/2017]

Les identifiants numériques et leur implémentation

ARCHER, Phil, GOEDERTIER, Stijn, LOUTAS, Nikolaos. Interoperability Solutions for European Public Administrations. *D7.1.3 - Study on persistent URIs, with identification of best practices and recommendations on the topic for the MSs*

and the EC. ISA, 2012. [en ligne] Disponible sur https://joinup.ec.europa.eu/sites/default/files/D7.1.3%20-%20Study%20on%20persistent%20URIs_0.pdf [consulté le 19/05/2017]

ARCHIMBAUD, Jean-Luc. *Identifiants des documents numériques : ISBN, ISSN, URL, Handle, DOI, OpenURL…* 2015. [en ligne] Disponible sur http://archivesic.ccsd.cnrs.fr/sic_01068135/ [consulté le 29/11/2016]

AYERS, Danny, VÖLKEL, Max. *Cool URIs for the Semantic Web*. W3C Interest Group Note, 2008. 16 p. [en ligne] Disponible sur https://www.w3.org/TR/cooluris/ [consulté le 29/11/2016]

BERNERS-LEE, Tim. *Cool URIs don't change.* W3C, 1998. [en ligne] Disponible sur https://www.w3.org/Provider/Style/URI [consulté le 31/03/2017]

BERNERS-LEE, Tim. *What do HTTP URIs Identify?* W3C, 2002. [en ligne] Disponible sur https://www.w3.org/DesignIssues/HTTP-URI.html [consulté le 31/03/2017]

BERNERS-LEE, Tim. *Uniform Resource Identifier (URI): Generic Syntax.* The Internet Society, 2005. [en ligne] Disponible sur http://www.ietf.org/rfc/rfc3986.txt [consulté le 12/04/2017]

BOOTH, David. *Four Uses of a URL : Name, Concept, Web Location and Document Instance*. W3C, 2003. [en ligne] Disponible sur https://www.w3.org/2002/11/dbooth-names/dbooth-names_clean.htm [consulté le 31/03/2017]

ERPA Seminar. *Persistent Identifiers, Final Report*. Cork, Ireland. 17-18 June 2004. Séminaire Erpanet, [en ligne] Disponible sur http://www.erpanet.org/events/2004/cork/Cork%20Report.pdf [consulté le 12/04/2017]

GREEN, Brian, BIDE, Mark. *Unique Identifiers: a brief introduction*. 1999. 11p. [en ligne] Disponible sur http://www.bic.org.uk/files/pdfs/uniquid.pdf [consulté le 29/11/2016]

HILSE, Hans-Werner, KOTHE, Jochen. *Implementing Persistent Identifiers*. Consortium of European Research Libraries, 2006. Disponible sur: http://webdoc.sub.gwdg.de/edoc/ah/2006/hilse_kothe/urn%3Anbn%3Ade%3Agbv%3A7-isbn-90-6984-508-3-8.pdf [consulté le 12/04/2017]

KUNZE, John A. *Towards Electronic Persistence Using ARK Identifiers*. California Digital Library, 2003. [en ligne] Disponible sur https://wiki.ucop.edu/download/attachments/16744455/arkcdl.pdf [consulté le 19/05/2017]

Ministère de la Culture et de la Communication. *Identifiants pérennes pour les ressources culturelles ; Vade-mecum pour les producteurs de données*. Version 1.0. 2015. [en ligne] www.bnf.fr/documents/identifiants_perennes_vademecum.pdf [consulté le 29/11/2016]

PASKIN, Norman, RUST, Godfrey. *Principles of identification*. Linked Content Coalition, 2014. [en ligne] Disponible sur http://www.linkedcontentcoalition.org/phocadownload/principles_of_identification/LCC%20Principles%20of%20Identification%20v1.1.pdf [consulté le 29/11/2016]

PASKIN, Norman. *Towards a rights data dictionary, Identifiers and semantic at work on the net*. EPS, 2002. [en ligne] https://www.doi.org/topics/020522IMI.pdf [consulté le 29/11/2016]

SOWA, John. Re : [ontolog-forum] OWL and lack of identifiers. Ontolog-Forum, 2007. [en ligne] Disponible sur http://ontolog.cim3.net/forum/ontolog-forum/2007-04/msg00030.html [consulté le 09/08/2017]

THOMPSON, Henry S., ORCHARD, David. *URNs, Namespaces and Registries*. W3C, 2006. [en ligne] Disponible sur https://www.w3.org/2001/tag/doc/URNsAndRegistries-50 [consulté le 31/03/2017]

WOOD, David. *Linking Enterprise Data*. Springer, 2010. Chapitre "Reliable and Persistent Identification of Linked Data Elements", p149-177.

Le Web et le document

BERNERS-LEE, Tim. *On the Next Web*. TED Talks, 2009. Captation vidéo de la conférence. [en ligne] Disponible sur http://www.ted.com/talks/tim_berners_lee_on_the_next_web [consulté le 31/03/2017]

BOCKAERT, Joël. *La communication du vivant : de la bactérie à Internet*. O. Jackob, 2017. 205p.

BOULTON, Jim. *The idea of the internet was born in Belgium*. Digital-archaeology.org [en ligne] Disponible sur http://digital-archaeology.org/the-idea-of-the-internet-was-born-in-belgium/ [consulté le 11/08/2017]

DAVIS, Ian. *Is 303 really necessary?* Blog de Ian Davis, 2010. [en ligne] Disponible sur http://blog.iandavis.com/2010/11/04/is-303-really-necessary/ [consulté le 01/06/2017]

ERTZSCHEID, Olivier. *De quoi la page Web est-elle le nom ? L'enluminure du code*. Blog Affordances, 2011. [en ligne] Disponible sur http://affordance.typepad.com/mon_weblog/2011/03/de-quoi-page-web-est-le-nom.html [consulté le 29/11/2016]

JACOBS, Ian, WALSH, Norman. *Architecture of the World Wide Web, Volume One*. W3C Recommendation, 2004. [en ligne] Disponible sur https://www.w3.org/TR/webarch/ [consulté le 09/08/2017]

KAHN Robert, E, LYONS, Patrice A. Representing Value as Digital Objects, a Discussion of Transferability and Anonymity. *D-Lib Magazine*, 2001. Vol.7, n°5. ISSN 1082-9873. DOI 10.1045/may2001-kahn

LATOUR, Bruno. Dans JACOB, Christian *(dir.). Lieux de savoir, Espaces et communautés*. Albin Michel, 2007. Chap. Pensée retenue, pensée distribuée. p605-615

LEGG, Catherine. *Pragmaticsm on the semantic web*. 2010. [en ligne] Disponible sur http://www.nordprag.org/nsp/1/Legg.pdf [consulté le 01/06/2017]

LICKLIDER, Joseph. Man-computer symbiosis, *IRE Transactions on Human Factors in Electronics*, Vol. HFE-1, Laboratoire d'informatique et d'intelligence artificielle MIT, 1960. Pp 4-11

Ministère de la Culture et de la Communication. *Métadonnées culturelles et transition Web 3.0, Feuille de route stratégique*. 2014. [en ligne] Disponible sur https://www.inha.fr/_attachments/le-web-semantique-pour-les-donnees-culturelles-actualite/64776-feuille-de-route-strategique-metadonnees-culturelles-et-transition-web-3-0%25282%2529.pdf?download=true [consulté le 22/03/2017]

RIFKIN, Jérémy. *L'âge de l'accès, la nouvelle culture du capitalisme*. La Découverte, 2005. 406p.

SALAÜN, Jean-Michel. *Les sept piliers de l'économie du document.* 2006. [en ligne] Disponible sur http://blogues.ebsi.umontreal.ca/jms/index.php/post/2006/10/05/86-les-sept-piliers-de-l-economie-du-document [consulté le 18/08/2017]

SALAÜN, Jean-Michel. *Vu Lu Su ; les architectes de l'information face à l'oligopole du web*. Paris : Éditions La Découverte, 2012. Chapitre 3, Réingénierie documentaire, p. 63-91.

STOCKWELL, Foster. *A history of Information storage and retrieval*. Mc Farland&Co, Jefferson, 2007. 208 p.

W3C. *Common http Implementation Problems*. W3C, 2003. [en ligne] Disponible sur https://www.w3.org/TR/chips/ [consulté le 31/03/2017]

WRIGHT, Alex. *Exploring a 'Deep Web' that Google can't grasp*. www.nytimes.com, 2009. [en ligne] Disponible sur http://www.nytimes.com/2009/02/23/technology/internet/23search.html [consulté le 13/07/2017]

Etudes de cas spécifiques

CAMPBELL, Debbie. *Reading locally, learning globally : creating universal experience – a national library view*. NLA, 2010. [en ligne] Disponible sur http://www.nla.gov.au/content/reading-locally-learning-globally-creating-a-universal-experience-a-national-library-view [consulté le 10/08/2017]

DALBIN, Sylvie, BERMES, Emmanuelle, ISAAC, Antoine, WENZ, Romain, NICOLAS, Yann, MERABTI, Tayeb, ANGJELI, Anila, FRANCART, Thomas, ROZAT, Lise, VANDENBUSSCHE, Pierre-Yves, VATANT, Bernard, RAIMOND,

Yves, COTTE, Dominique. Approches documentaires : priorité aux contenus. *Documentaliste-Sciences de l'Information*. ADBS, 2011. N° 48. p42-59

DAVIDSON, Paul, CIO Sedgemoor, District Council. *Designing URI Sets for the UK Public Sector*. Chief Technology Officer Council, 2009. [en ligne] Disponible sur https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/60975/designing-URI-sets-uk-public-sector.pdf [consulté le 18/05/2017]

FRANCART, Thomas. *Etude de cas, Office des Publications de l'Union Européenne*. [en ligne] Disponible sur http://www.sparna.fr/reference/office-des-publications-de-lunion-europeenne/ [consulté le 08/08/2017]

GARRITY, George M., THOMPSON Lorraine M., USSERY, Dave W., PASKIN, Norman, BAKER, Dwight, DESMETH, Philippe, SCHINDEL, David E., ONG, Perry S. 7th meeting. *Study on the identification, tracking and monitoring of genetic resources*. UNEP/CBD/WG-ABS/7/INF/2, 2009. 98 p. [en ligne] Disponible sur https://www.cbd.int/doc/meetings/abs/abswg-08/information/abswg-08-abswg-07-inf-02-en.pdf [consulté le 29/11/2016]

HOLLEY, Rose. *Resource sharing in Australia : 'Find' and 'get' in trove – making 'getting' better*. NLA, 2011. [en ligne] Disponible sur https://www.nla.gov.au/content/resource-sharing-in-australia-find-and-get-in-trove-making-getting-better [consulté le 10/08/2017]

SCHMITZ, Peter. Common Access to EU Information based on semantic technology. *The Multilingual Web – The Way Ahead*. W3C Workshop, 2012. [en ligne] Disponible sur https://www.w3.org/International/multilingualweb/luxembourg/slides/41-schmitz.pdf [consulté le 08/08/2017]

The National Library of Australia, *Persistent Identification Systems, Report on a consultancy conducted by Diana Dack for the NLA*. 2001. Disponible sur: http://www.imaginar.org/taller/dppd/DPPD/105%20pp%20Persistent.pdf [consulté le 12/04/2017]

BnF

BERMES, Emmanuelle. *Des identifiants pérennes pour les ressources numériques, l'expérience de la BnF.* Bibliothèque nationale de France, 2006. 9p.

BnF. *Semantic Web and data model*. Data.bnf.fr, 2017. [en ligne] Disponible sur http://data.bnf.fr/en/semanticweb#Ancre3 [consulté le 19/05/2017]

PEYRARD, Sébastien, TRAMONI, Jean-Philippe, KUNZE, John. The ARK Identifier Scheme : Lessons Learnt at the BnF and Questions Yet Unanswered. DC-2014 *Metadata Intersections: Bridging the Archipelago of Cultural Memory*. International Conference on Dublin Core and Metadata Applications, USA, 2014. [en ligne] Disponible sur http://dcevents.dublincore.org/IntConf/dc-2014/paper/view/241

WENZ, Romain. *Data.bnf.fr : describing library resources through an information hub*. Semantic Web in bibliotheken, 2010. Captation vidéo de la conférence. [en ligne] Disponible sur http://www.scivee.tv/node/27093 [consulté le 31/03/2017]

WENZ, Romain, SIMON, Agnès. « Data.bnf.fr: FRBR and Linked Data at the French National Library ». *Scatnews*, p.3-4. IFLA, 2011 [en ligne]. Disponible sur : http://www.ifla.org/files/cataloguing/scatn/scat-news-36.pdf [consulté le 04/05/2017]

WENZ, Romain. « Web sémantique, Open Data et bibliothèques : l'exemple de data.bnf.fr ». *Le Blog du Labo BnF*. BnF, 2012 [en ligne] Disponible sur : http://labobnf.blogspot.fr/2012/09/web-semantique-open-data-et.html [consulté le 04/05/2017]

WENZ, Romain. « Data.bnf.fr : au-delà des silos ». Dans Approches documentaires : priorité aux contenus. *Documentaliste-Sciences de l'Information*. ADBS, 2011. N° 48. p42-59

BBC

KOBILAROV, Georgi, SCOTT, Tom, RAIMOND, Yves, OLIVER, Silver, SIZEMORE, Chris, SMETHURST, Michael, BIZER, Christian, LEE, Robert. *Media Meets Semantic Web – How the BBC Uses DBpedia and Linked Data to Make Connections*. British Broadcasting Corporation, Londres. Freie Universität, Berlin. Rattle Research, Sheffield. 2009. [en ligne] Disponible sur https://pdfs.semanticscholar.org/5728/393384a3a60e55a72fa3a01d4fc1b258aac2.pdf [consulté le 01/06/2017]

RAIMOND, Yves. « Les programmes de la BBC tirent avantage du web de données ». Dans Approches documentaires : priorité aux contenus. *Documentaliste-Sciences de l'Information*. ADBS, 2011. N° 48. p57

RAIMOND, Yves, FERNE, Tristant, SMETHURST, Michael, ADAMS, Gareth. *The BBC World Service Archive Prototype*. BBC R&D, 2014. [en ligne] Disponible sur www.sciencedirect.com/science/article/pii/S1570826814000535 *[consulté le 01/06/2017]*

SINCLAIR, Patrick, HUMFREY, Nicholas, RAIMOND, Yves, SMETHURST, Michael, SCOTT, Tom. *The Web as a Content Management System*. BBC Audio and Music Interactive, 2009. [en ligne] Disponible sur http://www2009.eprints.org/236/1/www2009developers_submission_51.pdf

RAIMOND, Yves, SCOTT, Tom, OLIVER, Silver, SINCLAIR, Patrick, SMETHURST, Michael. Dans WOOD, David. *Linking Enterprise Data*. . Springer, 2010. Chap: Use of Semantic Web technologies on the BBC Web Sites, p263-283

ANNEXES

Table des annexes

TABLEAU COMPARATIF DES IDENTIFIANTS

Légende :

	Information non trouvée
N/A	Information non disponible/non applicable/non accessible

Acronyme	Nom complet	Origine	Type	Structure régente	Structure régente en France	Année de naissance	Norme	Fonction(s)	Ressources concernées	Lien avec d'autres identifiants	Nombre de caractères	Payant/Gratuit	Exemple
ISBN	International Standard Book Number	Publisher Association	Identifiant ISO	International ISBN Agency	Bibliothèque Nationale de France	1972	ISO 2108:2005	Singularisation	Livres et publications	EAN	13	Payant pour l'attribution d'un préfixe d'éditeur	979 10 000001 1 5
ISBN-A	International Standard Book Number Actionnable		Identifiant ISO	DOI, International ISBN Agency	DOI			Contextualisation et localisation	Livres et publications	DOI	15		10.978.0000001/0000018
ISSN	International Standard Serial Number	UNESCO	Identifiant ISO	Centre International de l'ISSN (CIEPS)	Centre International de l'ISSN (CIEPS)	1975	ISO 3297:2007	Singularisation	Publications sérielles	EAN	8	Gratuit	0029-4713
ISSN-L	International Standard Serial Number -Link		Identifiant ISO	Centre International de l'ISSN (CIEPS)	Centre International de l'ISSN (CIEPS)			Singularisation et contextualisation	Publications sérielles	ISSN	8		ISSN-L 0029-4714
ISTC	International Standard Text Code		Identifiant ISO	International ISTC Agency	ISTC Agency Cercle de la librarie Electre	2009	ISO 21047 :2009	Singularisation et localisation	Contenus textuels divers	ISBN	16	Gratuit	ISTC 0A9 2001 12B4A109 9
ISMN	International Standard Music Number		Identifiant ISO	International ISMN Agency	Société des Editeurs et Auteurs de Musique (SEAM)	1993	ISO 10957	Singularisation	Partitions et publications musicales écrites	EAN	13		979-0-2586-2569-6
ISWC	International Standard musical Work Code	Confédération Internationale des Sociétés d'Auteurs et Compositeurs (CISAC)	Identifiant ISO	International ISWC Agency	Société des Auteurs, Compositeurs et Editeurs de Musique (SACEM)	2001	ISO 15707:2001	Singularisation	Compositions musicales	IPI	11		ISWC T-034 698 521-1
ISRC	International Standard Recording Code	International Federation of Phonographic Industry (IFPI)	Identifiant ISO	International Federation of Phonographic Industry (IFPI)	Société Civile des Producteurs de Phonogrammes en France (SPPF) et Société Civile des producteurs Phonographiques (SCPP)	1986	ISO 3901 :2001	Singularisation	Enregistrement d'une œuvre musicale	N/A	12	Gratuit	ISRC FR - Z03 – 98 - 00212
ISAN	International Standard Audiovisual Number	Organisations professionnelles du secteur du cinéma et de l'audiovisuel	Identifiant ISO	Agence Internationale ISAN (ISAN-IA)	Agence française ISAN	2002	ISO 15706:2007	Singularisation	Publications audiovisuelles	N/A	24	Payant pour chaque ressource identifiée	ISAN0000-3BAB-0000-0000-G-0000-0000-Q
DOI	Digital Object Identifier	Association of Amercian Publishers	Identifiant ISO	Internation DOI Foundation (IDF)	Institution de l'Information Scientifique et Technique (INIST) CNRS	Normalisé ISO en 2012	ISO 26324	Singularisation et localisation	Tous types de ressources	Tous types d'identifiants : ISSN, ISBN…	Variable	Payant pour chaque ressource identifiée	10.1000/123456
ISCI	Identifiant international Normalisé des Collections	Helsinki University Library	Identifiant ISO	Danish Agency for Culture and Palaces	Agence Bibliographique de l'Enseignement Supérieur (ABES)	Normalisé ISO en 2012	NF ISO 27730	Singularisation et contextualisation	Collections, fonds et séries archivistiques	ISIL	Variable, jusqu'à 12 caractères		FI-H:Fennica
ISLI	International Standard Link Identifier	Projet d'association entre un texte et leur version en audiodescription	Identifiant ISO	International Information Content Industry Association (ICIA)		Normalisé en 2015	ISO 17316	Singularisation et localisation	Liens (Source-cible-relation)	N/A	Variable	Gratuit	ISLI 116063-4520086293791473426443001-9
ISNI	International Standard Name Identifier		Identifiant ISO	ISNI International Agency	Bibliothèque Nationale de France	2012	ISO 27729:2012	Singularisation et localisation	Personnes et organismes	VIAF, ORCID	16	Gratuit	0000 0001 2143 0518
ISIL	International Standard Identifier for Libraries and Related Organizations		Identifiant ISO	Danish Agency for Culture and Palaces	Agence Bibliographique de l'Enseignement Supérieur (ABES)	Normalisé en 2003	ISO/DIS 15511	Singularisation	Institutions culturelles: bibliothèques et services d'archives	N/A	Variable, jusqu'à 16 caractères	Gratuit	CA-QMCB

UUID	Universally Unique Identifier	Open Software Foundation (OSF)	Identifiant ISO	Non géré uniformément	Non géré uniformément	RFC en 2005 et ISO en 2008	ISO/IEC 9834-8 :2008	Singularisation	Tous types de ressources, mais surtout les composants logiciels	N/A	32	Gratuit	df0ad42d-fbeb-44ea-81a7-6c14b83f1e53
ARK	Archival Resource Key	California Digital Library (CDL)	Global issu d'initiative individuelle	California Digital Library (CDL)	Bibliothèque Nationale de France	2004	Non normé	Singularisation et localisation	Tous types de ressources	Identifiants d'initiatives individuelles	Variable	Gratuit	ark 12148/000000000000/version1
BICI	Book Item Contribution Identifier	Book Industry Communication	Global issu d'initiative individuelle			1997	Draft NISO en 2000 non abouti	Singularisation	Contenus textuels divers	SICI (deprecated), ISBN	Variable		BICI: 0521416205(1993)(10:EAAWL:234-261)2.2.TX;1-1
ELI	European Legislation Identifier	Conseil Européen	Global issu d'initiative individuelle	Office des publications de l'Union Européenne	Office des publications de l'Union Européenne	2011	Non normé	Singularisation et contextualisation	Textes de loi européens	NOR	Variable	Gratuit	eli/decret/2017/8/25/INTD 1713523D/jo/texte
HDL	Handle	Defense Advanced Research Projects Agency	Global issu d'initiative individuelle	DONA Foundation, Corporation for National Research Initiatives (CNRI)	DONA Foundation, Corporation for National Research Initiatives (CNRI)	1994	Non normé	Singularisation et localisation	Tous types de ressources	DOI	Variable	Payant pour l'attribution d'un préfixe d'éditeur	HDL : 12345/4561
ORCID	Open Researcher and Contributor Identifer	ORCID	Global issu d'initiative individuelle	ORCID	ORCID	2012	Conforme au standard ISO 27729:2012 (ISNI)	Singularisation	Chercheurs	ISNI	16	Gratuit	0000-0001-9873-1538
IPI	Interested Party Information	Confédération Internationale des Sociétés d'Auteurs et Compositeurs (CISAC)	Global issu d'initiative individuelle	Société Suisse pour les Droits des Auteurs d'Œuvres Musicales (SUISA)	Société Suisse pour les Droits des Auteurs d'Œuvres Musicales (SUISA)	2001	Non normé	Singularisation	Artistes et créateurs de contenus musicaux	ISWC	11	Attribution lors d'une adhésion à une Société de gestion des droits d'auteurs	I-000000000-0
EIDR	Entertainment Identifier Registry	Entertainment Identifier Registry Association (EIDRA)	Global issu d'initiative individuelle	Entertainment Identifier Registry Association (EIDRA)	Entertainment Identifier Registry Association (EIDRA)	2010	Non normé	Singularisation	Publications audiovisuelles	DOI, ISAN	21 (+ le préfixe DOI)	Payant pour l'adhésion à l'Association ou via l'enregistrement par des services	10.5240/0000-0000-0000-0000-0000-C
LSID	Life Science Identifier	Interoperable Informatics Infrastructure Consortium (I3E)	Global issu d'initiative individuelle	Object Management Group (OMG)	Object Management Group (OMG)		Non normé	Singularisation et localisation	Elements biologiques et scientifiques	Identifiants d'initiatives individuelles	Variable	Gratuit	lsid :ebi.ac.uk :SWISSPROT. accession :P34355 :3
PII	Publisher Item Identifier	Scientific and Technical Information publishers (STI Group)	Identifiant local	Scientific and Technical Information publishers (STI Group)	Scientific and Technical Information publishers (STI Group)	1995	Non normé	Singularisation	Articles en pré-publication	ISBN, ISSN	17	N/A	S0960-9822(11)01319-4
IMDB	Internet Movie Database	Internet Movie Database	Identifiant local	Internet Movie Database	Internet Movie Database	1990	Non normé	Singularisation et localisation	Productions audiovisuelles et contributeurs	N/A		N/A	www.imdb.com/title/tt123456/
TAG URI	Tag URI	Sandro Hawke, Tim Kindberg	Identifiant local	Non géré uniformément	Non géré uniformément	2001	Non normé	Singularisation et localisation	Tous types de ressources	N/A	Variable	N/A	tag :maurice.dupond@aol.fr,2017-05-05,00000001
OAI	Open Archive Identifier	Open Archive Initiative (OAI)	Identifiant local	Open Archive Initiative (OAI)	Open Archive Initiative (OAI)		Non normé	Singularisation et localisation	Tous types de ressources	Identifiants d'initiatives individuelles	Variable	N/A	oai : arXiv.org : 0000000
PPN	Pica Production Number	Stichting Pica Foundation, Online Computer Library Center (OCLC)	Identifiant local	Online Computer Library Center (OCLC)	Online Computer Library Center (OCLC)	1999	Non normé	Singularisation et localisation	Notices bibliographiques	N/A	9	Frais liés à l'implémentation de la solution PICA	PPN 142914614
HALid	Identifiant HAL	Centre pour la Communication Scientifique Directe (CCSD)	Identifiant local	Centre pour la Communication Scientifique Directe (CCSD)	Centre pour la Communication Scientifique Directe (CCSD)	2001	Non normé	Singularisation et localisation	Publications scientifiques	OAI, DOI, arXiv, PubMed, ADS	Variable	N/A	hal-123456789 – version 1

GLOSSAIRE

Actionnable (identifiant actionnable) : Un identifiant est actionnable s'il permet l'accès au référent qu'il identifie. Ceci est généralement possible lorsque l'identifiant possède en préfixe un protocole permettant le déréférencement sur le web (par exemple, les URI http sont typiquement des identifiants actionnables).

Affordance : L'affordance est la capacité d'un objet à suggérer sa provenance et/ou l'usage que l'on peut en avoir. Cette notion est liée à l'intuitivité et à la signifiance qui se dégage par exemple d'un identifiant.

Authoring : L'*authoring* correspond au processus de création d'une application, d'un site ou d'un programme qui comporte du texte, des sons, des vidéos etc. dans une logique hypertextuelle, permettant la navigation interne et externe à partir de l'objet en question.[439]

Bien informationnel : Le bien informationnel est définit par 7 caractéristiques propres à lui seul [440]:

- La non-destruction. Il ne se détruit pas lors de la consommation,
- Le prototype. Tous les biens informationnels sont des prototypes,
- L'interprétation. Tout produit informationnel s'inscrit dans un processus de double interprétation, on ne sait pas à qui il s'adresse,
- La plasticité. Le bien informationnel est malléable et peut se reconstituer pour composer un autre bien,
- L'expérience. Le bien informationnel est une expérience pure. On ne peut pas revenir en arrière une fois qu'on l'a consommé,
- L'attention. Il ne peut pas se consommer en même temps que d'autres biens informationnels, cela amène à une économie dite *de l'attention*,
- La résonnance. Plus un bien est consommé, plus il est connu et plus il sera consommé, de façon exponentielle.

Bottom-up/top-down (logique) : La logique *bottom-up/top-down* consiste à appréhender une organisation sociale, fonctionnelle ou sociétale selon des flux décisionnels et/ou procéduraux d'un point de vue hiérarchique. Soit du bas vers le haut (les petites structures remontent les informations vers la grande qui chapeaute le tout : *bottom-up*), soit du haut vers le bas (la structure centrale impose ses normes aux structures sous sa coupe et les influence : *top-down*).

Code Statut http : Les codes statuts sont des informations retournées au demandeur lors d'une requête http indiquant le résultat de l'interaction qui s'est produite. Les codes les plus courants sont :

- Code statut 200 : la requête a réussi,
- Code statut 303 : la requête est redirigée vers une autre URL,
- Code statut 404 : la page n'a pas été trouvée.

---

[439] *Authoring*. Businessdictionnary.com [en ligne] Disponible sur http://www.businessdictionary.com/definition/authoring.html [consulté le 25/08/2017]

[440] SALAÜN, Jean-Michel. *Les sept piliers de l'économie du document.* 2006. [en ligne] Disponible sur http://blogues.ebsi.umontreal.ca/jms/index.php/post/2006/10/05/86-les-sept-piliers-de-l-economie-du-document [consulté le 18/08/2017]

Collision : On nomme collision l'erreur informatique qui se produit lorsque qu'un même identifiant identifie deux ressources différentes, créant ainsi une ambigüité. Les risques de collision induisent que lorsque l'on rassemble deux bases d'identification en une seule, par exemple, certains identifiants se recoupent et ainsi créent de l'ambiguïté.

Core immutable ID, ou cœur immuable de l'identifiant : Le *core immutable ID* de l'identifiant correspond à une partie pérenne qui n'est jamais modifiée, qui est unique et qui en constitue la base immuable sur laquelle il est construit.

Cotation : La cotation est l'acte de coter, c'est-à-dire attribuer un identifiant (une cote) à un document dans le but de le nommer, de le localiser et/ou de le contextualiser. Celle-ci peut être déterminée par la place de l'objet dans une classification.

Crossmédia : Le cross-média correspond à un produit éditorial mis en scène dans un univers spécifique dont la richesse peut s'apprécier à travers différents médias : un livre complété par une bande audio, un film à la suite d'une Bande Dessinée, etc. Un bon exemple de cross-média est l'ensemble des produits dérivés, productions éditoriales, films, sites, jeux et œuvres qui découlent des univers de Tolkien ou de J.K Rowling, qu'elles soient de sources officielles (l'auteur ou le scénariste) ou sous forme de fan-art.

Crowdsourcing : Pratique correspondant à l'externalisation d'une activité habituellement réalisée par des professionnels d'un domaine vers « *un grand nombre d'acteurs anonymes (à priori).* »[441] Cela peut correspondre à des démarches folksonomiques d'attribution de métadonnées, par exemple.

Cybersquatting : Le *cybersquatting* est une pratique illégale consistant à faire enregistrer un nom de domaine pour soi reprenant explicitement le patronyme d'une marque en particulier dans le but de lui revendre par la suite, d'altérer son image ou de profiter de sa renommée.

Déréférencement : Le déréférencement est l'acte d'accéder à une ressource, une entité via son identifiant. Cela peut se faire au moyen d'un protocole, lorsque l'identifiant est actionnable, ou à travers des résolveurs.

Espace de noms : « Lieu » conceptuel désignant un ensemble de termes contenus dans une catégorie, de quel type que ce soit. Par exemple, « animal » est un espace de nom pouvant contenir les termes « chien, fourmilier, flamant rose, toucan… ».

Folksonomies : Pratiques correspondant au « tagging » en anglais, qui font référence à l'ajout de mots-clés libres par les utilisateurs qui vont « tagguer », appliquer des tags à des ressources suivant leurs propres appréhensions des sujets.

Follow your nose method : Méthode d'interopérabilité aussi nommée « navigation intuitive » consistant en la réutilisation des jeux de données existants.

Gold Standard Corpus : Le taux de fiabilité d'un système de NER (sa précision et son rappel) peut être mesuré en comparant ses résultats à des résultats obtenus par une reconnaissance humaine de termes. Ceux-ci sont des échantillons élaborés manuellement, nommés *Gold Standard Corpus**, et ils constituent l'idéal que l'on souhaite obtenir de manière automatique.

[441] BURGER-HELMCHEN, Thierry, PENIN, Julien. Crowdsourcing : définition, enjeux, typologie. *Management & Avenir*, Management Prospective Ed. 2011. N°41.

Hash/Hashage (méthode) : La mécanique de hashage, ou hash, (à bien distinguer des Hash URI ou dièse) est très associée aux systèmes de signature électronique. Il s'agit de récupérer une ressource, la faire absorber par un algorithme qui va coder chaque bit de celle-ci afin de produire un numéro complexe, qui sera complètement différent si le moindre octet était modifié dans le document par la suite. Ce « hash » sera en quelque sorte la carte d'identité unique de cette ressource telle qu'elle est exactement à un instant *t*.

Header : Littéralement en-tête, le *header* correspond à la partie en-tête d'un programme, d'une page, d'un élément codé informatique, souvent invisible à l'affichage, qui contient des informations structurelles et fonctionnelles.

Hive mind : Le concept de *hive mind*, littéralement « esprit de la ruche » est lié à la capacité qu'ont les insectes de réagir à l'unisson par un procédé de communication silencieux. Sur le web et notamment les réseaux sociaux, cela se traduit par l'expression d'une même idée au même moment par des personnes qui ne se connaissent pas entre elles, faisant ressortir ce que l'on pourrait définir comme une énorme entité à l'intelligence propre composée d'une pluralité d'entités plus petites.[442]

http range 14 : Le http range 14 est le nom d'une problématique informatique et des sciences de l'information qui concerne l'incapacité pour un demandeur de déréférencer un identifiant pointant sur une ressource abstraite.

*Hub and Spoke (*methode) : Egalement dénommé « roue et essieu », le *hub and spoke* est une méthode consistant à rassembler des informations dans un référentiel pour normer les termes et les URI, par exemple en créant des méta-référentiels tels qu'un thésaurus, un vocabulaire contrôlé, une liste d'autorité… Un référentiel qui serait au-dessus des autres et qui permettrait de les lier ensemble.

Hypertexte (technologie de) : La technologie de l'hypertexte correspond à l'ajout au sein d'un document numérique d'hyperliens, c'est-à-dire des liens cliquables permettant d'accéder à d'autres ressources. Cette interconnexion entre deux documents ne se fait que dans un seul sens, à la manière des citations.

Implémentation : L'implémentation est l'acte de mettre en place un système, l'interconnecter avec l'environnement et le rendre utilisable.

Interoperabilité : Le concept de l'interopérabilité correspond à « la capacité que possède un produit ou un système, dont les interfaces sont intégralement connues, à fonctionner avec d'autres produits ou systèmes existants ou futurs et ce sans restriction d'accès ou de mise en œuvre. »[443]

Link Maintenance Problem : Le problème de la maintenance des liens est lié à la difficulté de savoir où sont les liens cassés, à cause de l'uni-directionnalité des relations sur le web. Par exemple, si une ressource est supprimée ou modifiée, les personnes l'ayant citée sur leur site via un lien ne seront informés que lors du déréférencement.

Link rot : Processus par lequel les liens hypertextuels deviennent des liens morts, c'est-à-dire indisponibles de manière permanente. Il est lié à l'impossibilité

---

[442] *Hive mind*. Urbandictionnary.com [en ligne] Disponible sur http://www.urbandictionary.com/define.php?term=hivemind [consulté le 25/08/2017]

[443] *Définition de l'interopérabilité*. Groupe de travail Interopérabilité de l'AFUL. [en ligne] Disponible sur http://definition-interoperabilite.info/ [consulté le 18/08/2017]

de mettre à jour les liens lorsque les ressources changent (lié au concept du *link maintenance problem*).

Métadonnée : Donnée sur la donnée. Elle peut être de trois types [444]:

- Métadonnées de gestion : auteur, titre, date de création, date de modification…
- Métadonnées de description : sujet, description…
- Métadonnées de préservation : droits, format, source, résolution…

Négociation de contenu : Processus/mécanisme informatique par lequel un serveur est en mesure de proposer plusieurs représentations (par exemple, plusieurs formats ou plusieurs versions en langues différentes) d'un même document lorsque l'identifiant de celui-ci est déréférencé. Ce processus s'effectue généralement automatiquement de machine à machine en fonction des préférences spécifiées dans le *header* de la requête http.

Nom de domaine : Ils correspondent à une « adresse postale » sur Internet, un identifiant du domaine internet choisi. Les noms de domaine sont de plusieurs niveaux : le *top-level domain* (TLD) pouvant être les .com, .fr, .uk, etc. ; le *second-level domain* (SLD) pouvant préciser le TLD paris.fr, bnf.fr, etc. ou encore les sous-domaines www, tour-eiffel, etc.

Notice bibliographique : Fiche qui contient des éléments de description sur un document en vue d'être manipulée. Elle peut servir à des fonctions diverses : localisation, identification, contextualisation, description, etc.

OCR, océrisation : L'océrisation, ou OCR (*Optical Character Recognition*) correspond à des solutions logicielles reconnaissant automatiquement lors de la numérisation d'un texte papier les caractères présents, permettant ainsi un traitement informatique et une récupération des données (traitement de texte, copier-coller, etc.).

Ontologie : Répertoire classifiant les relations et déterminant les propriétés générales d'un ensemble d'objets.

Open-access : Littéralement « libre accès », concept issu du mouvement du même nom favorisant la mise à disposition en ligne des travaux des chercheurs et de manière plus générale de contenus en libre accès pour leur exploitation par tout un chacun.

Open-data : Littéralement « données ouvertes », concept issu du mouvement du même nom favorisant la mise à disposition en ligne de données liées aux activités d'une structure, notamment les organisations publiques, en vue de leur réutilisation et de leur exploitation par des tiers, dans une démarche de transparence économique, sociale et culturelle.

Open-source : Littéralement « code source ouvert » correspondant à la mise à disposition du code source d'un programme afin de favoriser son implémentation, sa réutilisation gratuite et l'ajout de modifications (améliorations par exemple) par des initiatives individuelles. Les logiciels *open-source* sont libres d'accès et fédèrent généralement une communauté active de tiers dans l'optique de son évolution constante.

PageRank : Nom de l'algorithme utilisé par Google pour classifier et référencer les pages moissonnées par son moteur de recherche. La place d'un site

[444] *Métadonnées*. Le Dictionnaire. Enssib. [en ligne] Disponible sur http://www.enssib.fr/le-dictionnaire/metadonnees [consulté le 18/08/2017]

dans le référencement de celui-ci en conditionne en général toute son utilisation, d'où les enjeux forts liés à ce système.

Peer-reviewing : Forme de validation par les pairs. Lorsqu'un article scientifique est soumis à publication il est lu et étudié par les confrères-consœurs spécialistes dont l'avis en détermine souvent la qualité.

Pensée Cartésienne : Pensée selon laquelle le sens d'un mot, d'un objet, d'une entité donnée n'est définit que par son concepteur, celui qui le pense.

Pensée Piercienne : Pensée selon laquelle le sens d'un mot, d'un objet, d'une entité donnée est complètement détaché de son propriétaire et évolue dans un « *cloud* » sémantique déterminé par les utilisations qui en sont faites.

Plug-in : Module d'extension d'un programme permettant d'ajouter des fonctionnalités non présentes sur le système de base. Ils sont normalement relativement facile à ajouter ou à enlever et constituent des petits systèmes enrichissant les possibilités initiales d'une utilisation.

Principe de l'opacité d'une URI : Dissociation consubstantielle qui existe systématiquement entre une URI et son contenu, peu importe le degré de signifiance de celle-ci.

Qualifiant/qualifieur : Petites chaînes de caractères ajoutées à la suite de l'identifiant (notamment employé pour les identifiants ARK) qui ne feront pas réellement partie de l'identification pérenne mais qui permettront entre autres : le *versioning* des entités, la gestion de services non liés à des contraintes de pérennité, la granularité des documents plus en finesse, et la définition des formats d'encodage.

Référent : Une entité devient le « référent » d'une identifiant lorsque celui-ci lui est attribué.

Référentiel : Un référentiel est un mode d'organisation des connaissances qui peut se décliner en plusieurs sous objets :

- Les ontologies,
- Les KOS (*Knowledge Organisation System*), tels que les thésaurus par exemple,
- Les jeux de données particulièrement importants qui deviennent des référentiels par la force des choses (notamment le nombre de réutilisations dont ils ont fait l'objet).

Résolveur : Système (programme ou plug-in) permettant de déréférencer un identifiant spécifique.

Scheme : Syntaxe ou partie de syntaxe d'un identifiant permettant son intégration dans un schéma connu (par exemple, on parle de *scheme* http).

Top-down : Voir *Bottom-up/Top-down* (logique).

Triplet (RDF) : Plus petite unité de donnée contenue dans un graphe RDF. Il constitue une « déclaration » composée de trois segments : un sujet, un prédicat, un objet. L'objet et le sujet sont liés par une propriété, le prédicat, qui donne une information sur le lien qui les unit. Tout le graphe RDF est composé de ces mini-phrases qui décrivent l'ensemble des liens présents entre les concepts définis en son sein.

Versioning : Le *versioning* est la capacité à gérer les différentes versions d'un objet, d'un document ou d'une entité en prenant en compte les différents stades de son évolution.

View Source Effect : Le *view source effect* est ce qui a conditionné la popularité du web, par le simple fait que tout le monde peut accéder au code source d'une page. Ainsi, cela a permis une compréhension et une appropriation rapide des individus et des sociétés de cette technologie et a beaucoup joué en faveur de son expansion.

TABLE DES MATIERES