

ÉCOLE NATIONALE DES CHARTES

Maxime Challon

licencié ès histoire

Les référentiels en institutions patrimoniales : évolution des pratiques et repositionnement

L'exemple des référentiels de l'Institut national de l'Audiovisuel

Mémoire pour le diplôme de master

« Technologies numériques appliquées à l'histoire »

2020

Résumé

Ce mémoire, réalisé pour l'obtention du diplôme de Master 2 « Technologies numériques appliquées à l'histoire » de l'École nationale des Chartes, retrace l'évolution des pratiques documentaires sur les référentiels en institution patrimoniale à travers l'étude des référentiels de l'Institut national de l'Audiovisuel (INA) et leurs alignements. Cette étude de l'évolution des formes et des structures des référentiels est liée à l'évolution de la place de ces référentiels au sein des systèmes documentaires, ainsi qu'aux besoins qui leur sont liés.

Mots-clés : institut national de l'audiovisuel ; référentiel ; thésaurus ; vocabulaire contrôlé ; vocabulaire hiérarchique ; ontologie ; web de données ; Wikidata ; liens ; alignement.

Informations bibliographiques : Maxime Challon, *Les référentiels en institutions patrimoniales : évolution des pratiques et repositionnement. L'exemple des référentiels de l'Institut National de l'Audiovisuel.*, mémoire de master « Technologies numériques appliquées à l'histoire », dir. Gautier Poupeau, École nationale des chartes, 2020.

Remerciements

Mes remerciements vont tout d'abord à Gautier POUPEAU, mon maître de stage, qui m'a accueilli, guidé, conseillé et intégré à son équipe malgré le travail à distance imposé par le contexte actuel. Je souhaite également remercier Axel ROCHE-DIORÉ pour ses explications et son soutien dans la réalisation technique de mon stage.

J'adresse aussi mes remerciements aux membres du pôle « Ingénierie de la Donnée », Lauryne LEMOSQUET, Otmane ELABBOUBI et Akli ABDI, ainsi qu'à Florence BRÉANT, cheffe de projet pour le *Lac de données*, pour le temps qu'ils m'ont accordé.

Que soit également remercié l'ensemble du département « Architecture et Innovation » de l'INA pour l'accompagnement fourni tout au long de mon stage, notamment Stanislas DE MAIGRET et Matthieu BORICAUD pour le déploiement de l'application, et Olivio SÉGURA pour la présentation des archives de l'INA.

Liste des abréviations

API Application Programming Interface

bibo the Bibliographic Ontology

CIDOC-CRM Comité International pour la Documentation - Conceptual Reference Model

DA Département des Archives Professionnelles

DDCOL Direction déléguée aux collections

DJ Direction juridique

DL Dépôt Légal

DSI Direction des systèmes d'information

EAD Encoded Archival Description

ÉPIC Établissement public à caractère industriel et commercial

FOAF Friend of a Friend

FRAD Functional Requirements for Authority Data

FRBR Functional Requirements for Bibliographic Records

FRSAD Functional Requirements for Subject Authority Data

GEMET General Multilingual Environmental Thesaurus

HTTP HyperText Transfer Protocol

INA Institut national de l'Audiovisuel

ISAN *International Standard Audiovisual Number*

ISNI *International Standard Name Identifier*

JSON JavaScript Object Notation

KOS Knowledge Organization Systems

LCSH Library of Congress Subject Headings

LED Linked Enterprise Data

LOD Linked Open Data

MARC MAchine-Readable Cataloging

- OAI-PMH** Open Archive Initiative Protocol for Metadata Harvesting
- OCLC** Online Computer Library Center
- ORTF** Office de la radio-télévision française
- OWL** Web Ontology Language
- RAMEAU** Répertoire d'autorité-matière encyclopédique et alphabétique unifié
- RDA** Resource Description and Access
- RDF** Resource Description Framework
- RDFS** RDF Schema
- SKOS** Simple Knowledge Organization System
- SPARQL** SPARQL Protocol and RDF Query Language
- UNIMARC** UNIversal MAchine-Readable Cataloging
- VIAF** Virtual International Authority File

Introduction

« Toutefois pour ne laisser cette quantité infinie ne la définissant point, [et] aussi pour ne jeter les curieux hors d'espérance et pouvoir acco[m]plir [et] venir à bout de cette belle entreprise, il me semble qu'il est à propos de faire comme les Médecins, qui ordonnent la quantité des drogues suivant la qualité d'icelles, [et] de dire que l'on ne peut manquer de recueillir tous ceux qui auront les qualitez [et] conditions requises pour estre mis dans une Bibliotheque.¹ »

EN 1627, Gabriel NAUDÉ compare le médecin au bibliothécaire, semblables par leur nécessité d'ordonner pour sélectionner, de classer pour retrouver, au milieu d'une masse d'objets. Cet ordonnancement, ce classement, passent pas une hiérarchisation de leur connaissance ou de leurs outils, dans le but de faciliter la recherche d'un médicament ou d'un livre pour l'utilisateur final. Cependant, plusieurs siècles plus tard, la hiérarchisation de la connaissance, ayant pour but de référencer une instance de la vie réelle, ne fonctionne plus : l'utilisateur ne part plus que très rarement d'un terme de la hiérarchie pour trouver son document ; il utilise le plus souvent un mot ou un concept qui le renverront vers une liste de résultats correspondant à sa requête. Alors, la notion de graphe prend le dessus sur celle de hiérarchie.

La notion évoquée de « quantité infinie » est aujourd’hui d'autant plus valable avec le web et l'explosion des quantités de données produites et stockées : avec cette mort de la notion de ressource, et par conséquent de celle de référentiel, la donnée structurée est implantée, peut être exploitée à la fois par une machine et par une personne, et est divisible et modulable à l'infini.

Cette transition de la ressource à la donnée, des référentiels hiérarchiques aux référentiels en graphe, est observable à l'INA. Créé en 1975 suite au démantèlement en sept sociétés de l'Office de la radio-télévision française (ORTF) par la loi du 07 août 1974,

1. Gabriel Naudé, *Advis pour dresser une bibliotheque. Tome 1 / . Presenté à monseigneur le president de Mesme. Par G. Naudé P....* T. 1, Chez François Targa, au premier pillier de la grand'salle du Palais, devant les Consultations, Paris, 1627, URL : <https://gallica.bnf.fr/ark:/12148/bpt6k1041429f> (visité le 01/09/2020), p.41-42.

l'INA est désigné comme un Établissement public à caractère industriel et commercial (ÉPIC) et « chargé de la conservation des archives, des recherches de créations audiovisuelles et de la formation professionnelle »². À ces missions est ajouté à partir de 1992 le dépôt légal de la télévision, de la radio, de la télévision satellite, par câble et numérique. Cette massification continue de documents et de données nécessite un classement et un référencement efficace des collections, ce qui a conduit à la création de plusieurs référentiels dans l'Institut.

Face à la croissance de l'utilisation du numérique, à l'accroissement des collections et des données à l'INA depuis les numérisations des collections au début des années 2000, aux nouveaux besoins exprimés par les professionnels et le public, une refonte du système documentaire est mise en place à la Direction des systèmes d'information (DSI) au sein du département « Architecture et Innovation » : les données et leurs métadonnées sont extraites des anciens silos de conservation, puis transformées et migrées dans un nouveau système d'information centralisé. Ainsi, les référentiels, descripteurs de chaque document, identificateurs de personnes ou d'instances des collections, subissent également ce traitement pour les uniformiser et permettre une homogénéisation et une meilleure valorisation des données de l'INA.

Cette migration massive permet d'observer l'évolution des pratiques documentaires de référencement et de description de ces dernières décennies, suivant la même évolution que l'ensemble du milieu bibliothéconomique en France, ainsi que les changements de structure des référentiels utilisés. La diversité de formes et de structures des référentiels montre que ces derniers sont considérés seulement comme des outils à disposition du documentaliste pour décrire ses fonds ; périphériques et éclatés, ils ne permettent pas une centralisation uniforme des données de l'INA.

Le projet du *Lac de données*, débuté en 2014, a pour but de centraliser l'ensemble des données de l'INA, les référentiels prenant alors une place centrale dans le nouveau système d'information. Ce projet s'inscrit dans l'évolution des besoins, tant chez les documentalistes que chez les utilisateurs, avec une utilisation désormais massive du web par tous les publics - chercheurs, professionnels des médias, jeunesse, ... - pour la recherche et la consultation de contenus. Cette éditorialisation croissante et indispensable nécessite de nombreuses données de référence, par lesquelles les contenus sont cherchables et trouvables.

2. *Loi n°74-696 du 7 août 1974 RELATIVE A LA RADIODIFFUSION ET TELEVISION*, 1974, URL : https://www.legifrance.gouv.fr/jo_pdf.do?id=JORFTEXT000000333539&pageCourante=08355 (visité le 01/09/2020), art.3.

Ce mémoire offre une réflexion sur ces évolutions des pratiques et des usages des référentiels à l'INA, et plus généralement dans une institution patrimoniale. Au-delà de ces évolutions sensibles, c'est le positionnement du référentiel au sein des systèmes documentaires qu'il est nécessaire d'interroger, de manière à faire face aux nouveaux enjeux et aux nouveaux besoins exprimés ces dernières années : d'un rôle périphérique, pensé comme un outil, le référentiel devient désormais un pivot autour duquel les données documentaires se raccrochent.

Mon stage, débuté en mai 2020 et terminé fin août 2020, à la DSI de l'INA, m'a permis d'intégrer le département « Architecture et Innovation » de Gautier POUPEAU, et plus particulièrement le pôle « Ingénierie de la Donnée » dirigé par Axel ROCHE-DIORÉ, afin d'effectuer une réflexion sur les méthodes d'alignement de plusieurs référentiels, et de mettre en œuvre ces méthodes. Les échanges avec mes collègues du pôle « Ingénierie de la Donnée » et les professionnels de la documentation de la Direction déléguée aux collections (DDCOL) et de la Direction juridique (DJ) m'ont permis de naviguer dans les référentiels, d'observer leurs différences, leurs structures, de comprendre les besoins qui leurs étaient associés ainsi que les difficultés impliquées par chaque référentiel dans l'opération d'alignement en vue de leur migration vers le *Lac de données*. Plusieurs missions m'ont ainsi été confiées :

- Extraire les fonctions et les occupations de personnes physiques depuis les notes qualité en texte libre du référentiel des personnes physiques et morales de la DDCOL, puis aligner ces fonctions extraites avec un thésaurus de noms communs propre à la DDCOL
- Aligner les personnes physiques de la DDCOL avec les entités correspondantes de Wikidata
- Aligner les fictions et les séries conservées à l'INA avec Wikidata de manière à récupérer également l'identifiant *International Standard Audiovisual Number* (ISAN)
- Aligner les référentiels de personnes physiques de la DJ et de la DDCOL, puis développer une interface de vérification et de complétion des alignements réalisés automatiquement

Ce mémoire retrace l'évolution des usages et des pratiques documentaires concernant les référentiels dans les institutions patrimoniales, en s'appuyant sur l'exemple des référentiels de l'INA. Dans un premier temps, dans une période allant jusqu'au début des années 2000, les référentiels sont uniquement considérés comme des fournisseurs de clés entre les données de manière à les contrôler plus facilement. Puis, jusqu'au milieu des années 2010, le web et le web de données permettent une mise en commun des référentiels qui se retrouvent alors liés entre eux. Enfin, depuis le milieu des années 2010, les

référentiels sont placés au centre des systèmes d'information : ils sont devenus les pivots des systèmes documentaires.

Première partie

**CONTRÔLER. A la recherche de
clés (années 1960 – fin des années
1990)**

Simple liste de mots ou thésaurus, un référentiel peut être conçu sous différentes formes. Dans un premier temps, son utilisation répond à un unique besoin : contrôler les formes que peuvent prendre les termes de manière à pouvoir les associer à plusieurs reprises à des documents, pour éviter une redondance de la même information à différents endroits. Le référentiel n'a alors qu'une fonction de clé. Cette opération, simple au premier abord, peut se complexifier avec l'intégration de synonymes ou de termes similaires associés au terme principal. L'arbre de classification, modèle du thésaurus, permet de contrôler un ensemble de termes tout en leur apportant du sens.

Les référentiels de l'INA reflètent cette évolution et les pratiques qui leur sont liées. L'INA est pourvu de listes de noms associés à des synonymes ou des variantes pour les noms de personnes, mais les noms communs — plus difficiles à exprimer de manière contrôlée — se trouvent eux dans un thésaurus permettant d'établir plus de relations sémantiques entre les termes similaires.

Chapitre 1

Le référentiel comme clé

Considéré comme une simple aide ou outil au service du documentaliste ou de l'utilisateur, le référentiel trouve d'abord sa place comme fournisseur de clés. Son utilisation principale est d'offrir au document décrit des vedettes qui puissent permettre une classification ou une recherche aisée de ce document. Cependant, pour être efficaces, ces vedettes doivent partager un langage contrôlé, des règles de graphie, de syntaxe, ...D'abord conservées sur des fichiers papier en institutions patrimoniales, ces vedettes ont été parmi les premiers éléments rétroconvertis, donnant naissance aux fichiers d'autorité numériques, et permettant une interopérabilité entre les référentiels par le biais des portails numériques.

1.1 Du langage libre au langage contrôlé : vers l'indexation

« La nature n'a pas juré de ne nous offrir que des objets exprimables par des formes simples de langage¹ »

Le langage permet aux hommes de communiquer entre eux. Ce langage libre, naturel, comprend l'ensemble des langues, et donne aux hommes la possibilité de décrire le plus précisément possible le monde qui les entoure, sans jamais atteindre la description idéale. Seulement, ce langage conduit à des variations graphiques ou syntaxiques, selon la déclinaison des noms ou la conjugaison des verbes ; la polysémie est également l'une des conséquences de ce langage naturel selon le contexte de chaque mot ; enfin, le langage libre conduit à la synonymie. Toutes ces caractéristiques du langage des hommes perturbe et complexifie la tâche de description documentaire, bien qu'elles soient essentielles à la communication entre eux.

1. Paul Valéry, *Variété III*, 9e éd, Paris, 1936, URL : <https://catalogue.bnf.fr/ark:/12148/cb41687051w>, p.18.

Afin de régler ces confusions possibles entre les mots et de régir leur formation, des langages contrôlés ont très vite fait leur apparition. Ils permettent de décrire des concepts, des thèmes, des ouvrages, tout en permettant un classement potentiel. Ce recours aux langages contrôlés est une pratique très ancienne, née avant l'apparition des *codices* lorsque déjà la recherche d'informations était nécessaire. Pratique millénaire, l'attribution de termes contrôlés à une information se perpétue encore actuellement, par exemple sous la forme de « hashtags » sur les réseaux sociaux, qui permettent de décrire un texte et de le retrouver ensuite aux côtés d'autres similaires.

Dans l'Antiquité, les index n'existent pas encore. Cependant, des vocabulaires contrôlés sont utilisés pour le classement et pour la mémorisation des textes. Ces termes contrôlés se retrouvent dans des notes marginales, des tables de concordance ou bien dans les catalogues. Au III^{ème} siècle av. J.C., Callimaque DE SILÈNE réalise le catalogue de la bibliothèque d'Alexandrie en utilisant le genre du texte pour lui déterminer une classe, puis les *volumina* sont rangés dans des rayons selon un ordre alphabétique, ces rayons reflétant les classes attribuées selon le genre.

Au Moyen-Âge, les premiers index apparaissent, s'ajoutant aux tables de concordance. Isidore DE SÉVILLE ne crée qu'un classement alphabétique dans son Livre X des *Étymologies*, sans indexé son ouvrage. Cinq siècles plus tard, les vedettes commencent à être normalisées dans certaines œuvres, le nominatif ou l'ablatif étant considérés comme la forme retenue, et rassemblées dans un index alphabétique².

Avec la Renaissance puis l'Ancien Régime, l'indexation devient plus fine et les index de fin de volumes sont de plus en plus imposants. Ils permettent au lecteur d'avoir un accès direct aux passages du texte contenant l'entrée d'index. Encore, ces index lient une

2. Jean BERGER, dans son analyse du *Liber de honoribus*, le plus vieil index alphabétique compilé au XII^{ème} siècle, étudie avec précision l'indexation des chartes du Cartulaire de Saint-Julien de Brioude : les lieux et les personnes sont ainsi indexés. Voir Jean Berger, “Indexation, memory, power and representations at the beginning of the 12th century : The rediscovery of pages from the tables to the “Liber de honoribus”, the first cartulary of the collegiate Church of St. Julian of auvergne (Brioude)”, *The Indexer*, 25–2 (oct. 2006), OCLC : 882418933, p. 95-99, URL : <http://halshs.archives-ouvertes.fr/halshs-00975166> (visité le 27/07/2020), pp.97 et suivantes

classification générale suivie d’alphabétique, tout en normalisant leurs entrées^{3 4}.

1.2 Une clé entre les données : les vocabulaires contrôlés

Dans les vocabulaires contrôlés, les termes servant à la description sont soumis à une normalisation. La maîtrise de la terminologie est l’objectif de ces vocabulaires ainsi que ce qui permet à ces derniers d’être une « colle qui tient l’ensemble du système⁵ » et le rend cohérent. Ces vocabulaires ne sont pas hiérarchisés et tirent la description de leur terme uniquement par leur graphie et leur désambiguïsation face au langage naturel. Ils permettent d’éviter les erreurs de graphie introduites par le documentaliste — par conséquent les différences de graphies — , d’éviter également les redondances de termes similaires et de rendre un système univoque. Ainsi, les vocabulaires contrôlés deviennent à eux seuls des langages propres à leurs utilisateurs⁶, servant à lutter contre la trop grande richesse du langage naturel humain. Pour effectuer le contrôle des termes, plusieurs points de contrôle sont introduits : le contrôle de la forme des vedettes, celui de la polysémie, et celui de la synonymie. L’exemple des autorités⁷ et des⁸, bien que comprenant une hiérarchie et des relations complexes, permettent d’observer la formation d’un langage contrôlé.

3. Jean-Daniel SCHOEPFLIN dans son *Alsatia illustrata* de 1751 créé ainsi deux index distincts : l’un pour les personnes (*Index auctorum*), l’autre pour les termes évoqués dans son œuvre(*Index rerum*). L’ensemble des noms est indexé au nominatif puis ils sont parfois subdivisés en thèmes ou événements. L’index devient ainsi indépendant de la graphie et de la grammaire de la langue utilisée. Voir Johann Daniel Schoepflin, *Alsatia illustrata Celtaica Romana Francica, auctor Jo. Daniel Schoepflinus, Consil. & Historiographus Regius, Histor. & Eloq. professor Argent. regiae inscriptionum ut & Anglic. Petropolit. Ac Corton Academiarum socius.* T. 2, 2 t., Ex typographia regia, Sumptibus Jo. Friderici Schoepflii, Colmariae [Colmar], 1751 (Collection Jacques Doucet), URL : <http://bibliotheque-numerique.inha.fr/idurl/1/12532> (visité le 26/07/2020). Voir Annexe A : Les index de la Renaissance, termes contrôlés et classification alphabétique (les index de l’*Alsatia Illustrata* de Jean-Daniel SCHOEPFLIN).

4. Robert ESTIENNE pousse plus loin encore l’indexation, un siècle et demi avant Jean-Daniel SCHOEPFLIN, en créant de multiples index : celui des populations, des villes, Ces index sont eux-mêmes subdivisés, normalisés et classés alphabétiquement, les rendant œuvre à part entière. Voir Robert Estienne, *Thesaurus linguae latinae, seu Promptuarium dictionum et loquendi formularum omnium ad latini sermonis perfectam notitiam assequendam pertinentium, ex optimis auctoribus concinnatum*, t. 2, 2 t., Country : FR, Lugduni, 1573, URL : <http://gallica.bnf.fr/ark:/12148/bpt6k8720517v> (visité le 27/07/2020).

5. « Controlled vocabularies have become the glue that holds the system together » in Louis Rosenfeld, Peter Morville et Jorge Arango, *Information architecture for the World Wide Web*, OCLC : 922954742, Beijing, 2015

6. Le Centre National de Ressources Textuelles et Lexicales (CNRTL) définit ainsi un vocabulaire : « Dictionnaire ne comportant que les mots les plus usuels d’une langue »

7. Bibliothèque nationale de France, *RAMEAU*, RAMEAU, URL : <https://rameau.bnf.fr/> (visité le 01/09/2020).

8. The Library of Congress, *Library of Congress Subject Headings*, URL : <https://id.loc.gov/authorities/subjects.html> (visité le 02/09/2020).

1.2.1 Contrôle de la forme des vedettes

La forme des vedettes doit être contrôlée de manière à offrir une graphie uniformisée ; plusieurs moyens sont alors utilisés :

- Choix d'un mot ou d'une locution en langage libre, le plus général possible, en évitant les ambiguïtés : le Répertoire d'autorité-matière encyclopédique et alphabétique unifié (RAMEAU) a fait le choix de « Télévision », de même que les Library of Congress Subject Headings (LCSH)
- Utilisation d'une langue définie pour l'ensemble du vocabulaire, sauf pour le cas d'emprunts : RAMEAU est en français, on y trouve alors la vedette « Droit d'auteur » au lieu de « Copyright », alors que les vedettes LCSH considèrent l'inverse : « Copyright » avec une variante en français renvoyant vers la vedette RAMEAU. Cependant, des variantes linguistiques sont attachées aux vedettes : l'italien « Televisione » est ainsi lié à la vedette « Télévision » de RAMEAU
- Utilisation majoritaire du pluriel pour les noms communs (comme la vedette RAMEAU « Livre ») ; le singulier étant utilisé pour les concepts généraux (« Écriture »)
- Choix d'une forme plus attestée ou plus usitée qu'une autre : nous pouvons trouver « Radiodiffusion » et non « Radio » dans RAMEAU ; de même, nous constatons la présence de « Radio broadcasting » dans LCSH, la vedette « Radio » étant réservée pour le moyen de communication

1.2.2 Contrôle de la polysémie et de l'homographie

L'ambiguïté du langage naturel dans la graphie et la polysémie peut induire le documentaliste et l'utilisateur en erreur, et réduire ainsi la puissance et l'utilité du vocabulaire mis en place. Contrôler la polysémie et l'homographie est par conséquent indispensable. Une vedette doit alors correspondre à un seul concept : deux actions sont alors possibles pour supprimer les ambiguïtés et améliorer le vocabulaire.

- L'ajout d'un qualificatif entre parenthèses peut permettre la levée de cette ambiguïté : RAMEAU utilise les qualificatifs « Plantes » et « Anatomie » pour traiter l'homonymie de « Iris » ; cette ambiguïté existant également en anglais, LCSH utilise les mêmes qualificatifs (« Plants » et « Eye »)
- L'utilisation de l'opposition singulier/pluriel permet de distinguer un concept abstrait d'une réalité concrète : RAMEAU utilise cette opposition de genre pour séparer le « Cinéma » compris comme art, du « cinéma » compris comme bâtiment où cet art est projeté

1.2.3 Contrôle de la synonymie

Le dernier écueil des vocabulaires contrôlés est la synonymie : source de confusions, il conduit à la création de nombreuses vedettes qui se rapportent finalement à un même concept. LCSH et RAMEAU ont fait le choix de créer des termes exclus qui renvoient vers le concept auquel ils sont reliés : ainsi, une recherche du terme « Détenus » dans RAMEAU renvoie vers la vedette « Prisonniers ». Les termes exclus peuvent être de différents types :

- des synonymes : « Cameramen », « Cinematographers », « Operating Cameraman » sont tous des termes exclus et synonymes de « Cameraman » dans les LCSH
- des abréviations ou des acronymes : l'abréviation « ISSN » est ainsi un terme exclu de l'*« International Standard Serial Numbers »* dans les LCSH
- des inversions de termes — qui permettent la mise en avant d'un terme important — : LCSH considère comme terme exclu de « Cameraman » « Operators, Camera »
- enfin, les termes exclus peuvent être des constructions syntaxiques, permettant de supprimer l'ambiguïté encore présente ou bien préciser le champ de la vedette : RAMEAU précise ainsi l'étendue géographique des vedettes en ajoutant le nom du pays après le concept ; la nouvelle vedette ainsi créée devient restrictive et spécifique. C'est le cas notamment de « Chaînes de télévision – France » qui précise la vedette « Chaînes de télévision ».

Ces termes exclus permettent de multiplier les points d'accès à un concept en prenant en compte la complexité du langage naturel qui désigne souvent par différents termes un même concept. Ainsi, deux utilisateurs cherchant la même vedette mais avec des termes différents pourront plus facilement retrouver cette vedette. Si ces termes ne sont pas obligatoirement des synonymes, leur contexte et le vocabulaire dans lesquels ils se trouvent les font se considérer comme synonymes⁹. Peter MORVILLE et Louis ROSENFELD nomment ces rapprochements des « Anneaux de synonymie »¹⁰ : ils connectent un ensemble de mots qui sont compris comme équivalents dans leur contexte d'utilisation¹¹.

1.3 Une clé entre les jeux de données : l'interopérabilité par les fichiers d'autorité et les portails

Comme nous l'avons évoqué précédemment (voir section 1.2 : Une clé entre les données : les vocabulaires contrôlés), les vocabulaires contrôlés sont de nouveaux langages,

9. L. Rosenfeld, P. Morville et J. Arango, *Information architecture for the World Wide Web...*

10. « Synonym rings » in *Ibid.* Voir Figure 1.1 : Anneau de synonymie du terme « Prisonniers » de RAMEAU et Figure 1.2 : Anneau de synonymie du terme « Prisoners » de LCSH.

11. « Connects a set of words that are defined as equivalent for the purposes of the retrieval. » in *Ibid.*

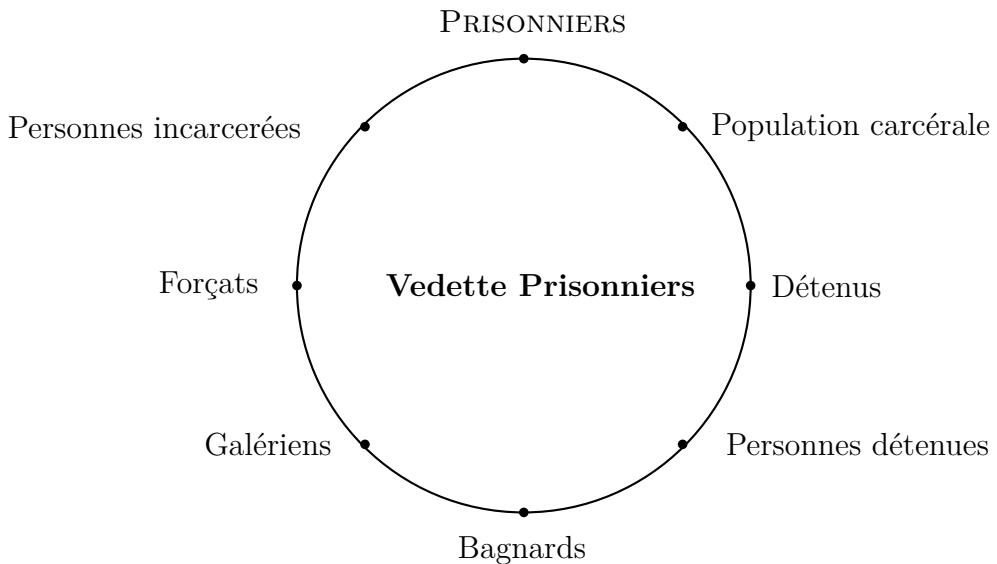


FIGURE 1.1 – Anneau de synonymie du terme « Prisonniers » de RAMEAU

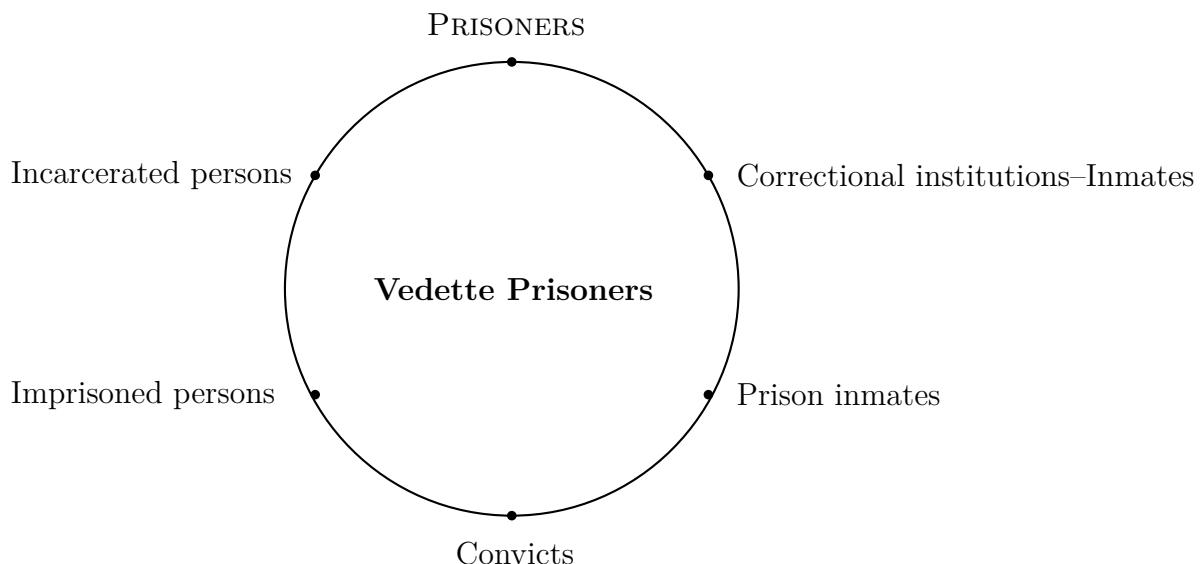


FIGURE 1.2 – Anneau de synonymie du terme « Prisoners » de LCSH

spécifiques et uniformisés, se substituant au langage naturel humain pour un domaine précis. Le vocabulaire est par conséquent un référentiel propre à l'institution qui l'a créée et a pour seul utilisateur cette institution. Seulement, deux institutions aux activités proches créent deux vocabulaires similaires, se distinguant par la complétude de certaines vedettes ou par des variantes de graphies.

Le domaine bibliothéconomique a été le premier à informatiser ses vocabulaires et ses fichiers d'autorités en masse, permettant ainsi une amélioration de l'expérience utilisateur et du catalogage, et un partage possible avec des institutions proches.

1.3.1 La naissance des autorités par rétroconversion

« Les fichiers d'autorité appartiennent bien à un ensemble : fonctionnant comme un tout, avec des règles d'interdépendance et d'interopérabilité de ses constituants, ils permettent le contrôle de la cohérence des métadonnées bibliographiques.¹² »

Avant la naissance du web, chaque ouvrage était décrit dans un catalogue et classé par ordre alphabétique des noms d'auteur. Des catalogues thématiques ont été créés, de même que des fichiers physiques en bibliothèque, permettant la recherche de documents selon un sujet précis. Cependant, l'indexation des documents est réduite au titre, à l'auteur, et à quelques sujets. En effet, la structure même d'un fichier papier en bibliothèque nécessite de dupliquer la notice d'un exemplaire en plusieurs notices qui vont être placées par la suite dans le fichier correspondant au sujet.

Ces fichiers physiques des bibliothèques, bien qu'utiles aux lecteurs par leur classement thématique, présentent plusieurs difficultés : d'abord, l'indexation se trouve limitée à quelques mots ; ensuite, la création d'un fichier thématique est complexe à réaliser par le choix des vedettes et produit alors un immense silence ; enfin, la consultation d'une fiche par un lecteur empêche un second de la consulter dans le même temps.

Dès les années 1970, les bibliothèques se sont engagées dans une vaste opération de rétroconversion de leurs notices documentaires. Les fichiers physiques et les notice cartonnées sont alors informatisés et « reproduits presque à l'identique [...] sous forme de bases de données »¹³. L'informatisation des notices et des fichiers permet par conséquent d'améliorer l'indexation des documents, et à l'utilisateur de pouvoir trouver plus de documents correspondant à sa recherche plus rapidement. Ainsi, les autorités LCSH, créées en 1914 sous format papier, ont été informatisées ; les autorités RAMEAU créées dans les années 1980 reprennent celles LCSH en les complétant.

Cependant, ces fichiers d'autorité comportent, comme nous l'avons évoqué plus haut (subsection 1.2.3 : Contrôle de la synonymie), des formes retenues et des formes rejetées des termes, ce qui crée de multiples renvois à l'intérieur du fichier physique ou informatique. L'arrivée des moteurs de recherche dans les années 2000 permet de supprimer ces différences de termes en indexant à la fois les formes retenues et les formes rejetées, permettant de trouver directement la vedette recherchée.

12. Vincent Boulet, François Mistral, Olivier Rousseaux, Yann Nicolas et Philippe Le Pape, *Arabesques n°85*, réd. par David Aymonin, t. 85, Montpellier, 2017 (Arabesques), URL : <http://www.abes.fr/Publications-Evenements/Arabesques/Arabesques-n-85> (visité le 14/07/2020), p.6.

13. [bermes_1_2013](#).

1.3.2 Partager des vocabulaires : à la recherche de la meilleure interopérabilité

La problématique du partage des référentiels entre institutions se pose avant l’informatisation des catalogues et des fichiers des bibliothèques. En effet, le format MAchine-Readable Cataloging (MARC), né en 1968 à la Bibliothèque du Congrès, permet l’échange de données entre les institutions et la « duplication des notices d’un catalogue à un autre »¹⁴. Malgré de multiples variantes nationales, l’UNIversal MAchine-Readable Cataloging (UNIMARC) reste aujourd’hui le format d’échange privilégié entre les bibliothèques.

Pour partager les fichiers d’autorité et aboutir à une interopérabilité totale des données entre deux institutions par le biais des machines, différents protocoles d’échange ont été utilisés — ou délaissés en fonction des difficultés imposées par chacun—. Dès les années 1980 est développé le protocole Z39-50. Ce protocole permet d’interroger une base de données de manière synchrone, selon la requête du client, et de récupérer des données en format MARC¹⁵.

Ce protocole Z39-50 est destiné aux catalogueurs qui peuvent ainsi « repérer puis télécharger une notice dans un catalogue distant plutôt que d’avoir à la saisir *ex nihilo* »¹⁶. Le partage, « par conversion et copie »¹⁷, n’est alors qu’une simple copie de données, dont la mise à jour est difficile. L’existence de ce protocole, bien que destiné aux professionnels de la documentation, a suscité la création de portails de consultation de notices documentaires ou de fichiers d’autorité, interrogeant de manière synchrone les bases de données : cette utilisation orientée utilisateur du protocole Z39-50 permet à la Bibliothèque nationale de France d’offrir différents services(intégration des notices dans Online Computer Library Center (OCLC), recherche dans le Catalogue Collectif de France (CCFR), ...¹⁸). Cependant, face aux temps de réponses importants et aux résultats appauvris retournés par la requête, les portails se sont révélés décevants et peu efficaces. De plus, l’utilisation d’un portail nécessite de la part de l’utilisateur qu’il connaisse précisément ce qu’il cherche de manière à se connecter au portail correspondant (sui lui-même doit être connu de cet utilisateur)¹⁹.

14. **bermes_2_2013.**

15. B. nationale de France, *Le protocole Z39.50*, BnF - Site institutionnel, URL : <https://www.bnf.fr/fr/le-protocole-z3950> (visité le 02/09/2020).

16. **bermes_2_2013.**

17. **bermes_2_2013.** Voir Annexe B : Les différents types d’interopérabilité

18. *Ibid.*

19. Sylvie Dalbin, Emmanuelle Bermès, Antoine Isaac, Romain Wenz, Y. Nicolas, Tayeb Merabti, Anila Angjeli, Thomas Francart, Lise Rozat, Pierre-Yves Vandebussche, *et al.*, “Approches documentaires : priorité aux contenus”, *Documentaliste-Sciences de l’Information*, Vol. 48-4 (2011), Publisher : A.D.B.S., p. 42-59, URL : <https://www-cairn-info.proxy.charter.psl.eu/revue-documentaliste-48-4-2011>

La multiplication des formats d'échanges — MARC et UNIMARC pour les bibliothèques, Encoded Archival Description (EAD) pour les archives —, ainsi que la volonté d'offrir au public lien entre les différentes bases de données patrimoniales, ont conduit à la création d'un nouveau protocole, Open Archive Initiative Protocol for Metadata Harvesting (OAI-PMH). Ce protocole asynchrone repose sur deux acteurs : le fournisseur qui met à disposition ses données dans un « entrepôt », et le moissonneur qui collecte ces données pour les intégrer à son système²⁰.

Cependant, si les performances du protocole sont améliorées avec OAI-PMH, les documents et les fichiers d'autorité ne peuvent pas être sélectionnés et filtrés : un format d'échange simple, minimal, est nécessaire. Ce format est le Dublin Core²¹ comprenant quinze champs d'informations. Ce partage de données et de métadonnées entre les institutions permet une « interopérabilité par le plus petit dénominateur commun »²², où ce dénominateur est le Dublin Core. Ce dénominateur commun peut néanmoins présenter un appauvrissement des données puisque les champs sont très réduits, ou au contraire permettre de grandes différences au sein d'un même champ.

Auteurs, catalogueurs et bibliothécaires ont très vite ressenti le besoin de se dégager du langage naturel de manière à renvoyer rapidement vers des passages de leur texte ou à décrire le plus précisément possible les documents, pour faciliter la lecture ou la recherche de l'utilisateur final. D'abord effectuées sur des supports papier, ces opérations de descriptions ont été informatisées et ont permis le partage de données et de métadonnées entre les institutions : les notices et les fichiers d'autorité disponibles sont la source constante d'interrogations quant au meilleur moyen de les mettre à disposition, tant pour le professionnel que pour le public. L'ouverture de ces vocabulaires a permis une amélioration des descriptions et une uniformisation des pratiques d'indexation.

Cependant, ces vocabulaires contrôlés restent peu précis et sont limités à leur terminologie pour en tirer le sens : il ne comprennent pas de terminologie sémantique, qui permettrait d'améliorer plus encore la description effectuée, en pouvant se référer aux termes parents, frères ou fils. L'anneau de synonymie évoqué (subsection 1.2.2 : Contrôle de la polysémie et de l'homographie) permet la prise en compte des synonymes, mais ne donne pas de sens supplémentaire à la vedette.

[sciences-de-l-information-2011-4-page-42.htm](http://www.bnfr.fr/sciences-de-l-information-2011-4-page-42.htm) (visité le 02/08/2020).

20. B. nationale de France, *Protocole OAI-PMH*, BnF - Site institutionnel, URL : <https://www.bnfr.fr/fr/protocole-oai-pmh> (visité le 02/09/2020).

21. *Dublin Core Metadata Initiative*, URL : <https://dublincore.org/specifications/dublin-core/dcmi-terms/> (visité le 02/09/2020).

22. **bermes_2_2013**. Voir Annexe B : Les différents types d'interopérabilité

Chapitre 2

L’arbre, un vocabulaire contrôlé hiérarchique

Nous l’avons évoqué (section 2.2 : Le *thésaurus*, vocabulaire contrôlé hiérarchique le plus fréquent), le contexte d’un terme de vocabulaire peut lui donner un sens complémentaire ou différent. La hiérarchisation des vocabulaires permet un ajout de contexte à chaque terme, mais également un accroissement de la précision de la définition donnée à ce terme. Le vocabulaire hiérarchique contrôlé le plus fréquent est le *thésaurus* : la diversité de ses relations et de ses caractéristiques lui permet une adaptation à chaque vocabulaire. Cependant, la hiérarchie n’offre plus assez d’autorités pour décrire précisément les données de l’INA.

2.1 L’arbre de Porphyre : origines et influences

La définition d’un terme est une réflexion millénaire, et la recherche d’un référentiel, d’un dictionnaire pur n’est toujours pas abouti — l’intelligence artificielle nécessitant des référentiels solides, la réflexion sur la pureté du dictionnaire utilisé est constante. Umberto Eco considère que le dictionnaire « ne devrait comporter, pour la définition d’un terme, que les propriétés nécessaires et suffisantes pour distinguer ce concept d’un autre »¹. Ces propriétés nécessaires à la définition du terme ne doivent pas être une connaissance du monde, mais bien des propriétés analytiques : « Animal » est une propriété analytique de « Chien » alors que l’aboiement est une connaissance.

La théorisation du dictionnaire remonte à l’Antiquité et a eu de nombreuses influences dans les systèmes classificatoires jusqu’à nos jours : les vocabulaires utilisés en institutions patrimoniales sont pour la plupart des hiérarchies de termes.

1. Umberto Eco, *De l’arbre au labyrinthe : [études historiques sur le signe et l’interprétation]*, trad. par Hélène Sauvage, 1 t., Paris, 2010, chap.1.

2.1.1 L'arbre de Porphyre

La pensée aristotélicienne considère la définition d'un terme comme la forme substantielle, c'est à dire les attributs essentiels : l'« homme » est un « Animal rationnel mortel »². L'assemblage de ces propriétés essentielles crée une définition, mais chacune de ces propriétés peut s'appliquer à d'autres entités.

Le commentateur des *Catégories* d'Aristote au III^{ème} siècle, Porphyre, établit des arbres pour décrire le monde : celui des « Substances » a le plus de postérité en étant « un ensemble hiérarchisé et fini de genres et de substances »³, partant du *Summus genus*, la Substance, pour atteindre une espèce indivisible, définie uniquement par ses attributs analytiques appelés genres⁴. Un arbre de Porphyre est par conséquent une succession de genres divisés en espèces qui deviennent elles-mêmes des genres.

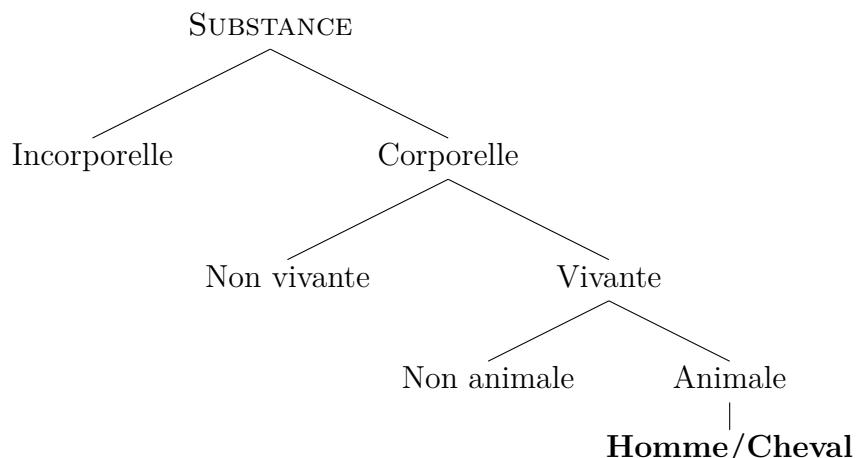


FIGURE 2.1 – Arbre porphyrien de l'homme avec les seuls attributs analytiques [d'après ECO (Umberto), *De l'arbre au labyrinthe : études historiques sur le signe et l'interprétation*], trad. par Hélène Sauvage, 1 t., Paris, 2010]

L'impossibilité de la distinction entre l'homme et le cheval impose de tenir compte des différences qui ne sont pas des attributs analytiques : « La rationalité est la différence de l'homme »⁵. Ainsi, ces différences vont s'ajouter aux genres des espèces. Ces différences deviennent elles-mêmes divisibles et constitutives : elles deviennent genre. Ces différences sont essentielles pour distinguer une espèce d'une autre (voir Figure 2.2 : Arbre porphyrien prenant en compte les différences).

Cependant, si la prise en compte des différences permet de différencier l'homme du cheval, elles ne permettent pas de distinguer le cheval de l'âne par exemple. Un même genre

2. *Ibid.*

3. *Ibid.*

4. Voir Figure 2.1 : Arbre porphyrien de l'homme avec les seuls attributs analytiques

5. *Ibid.*

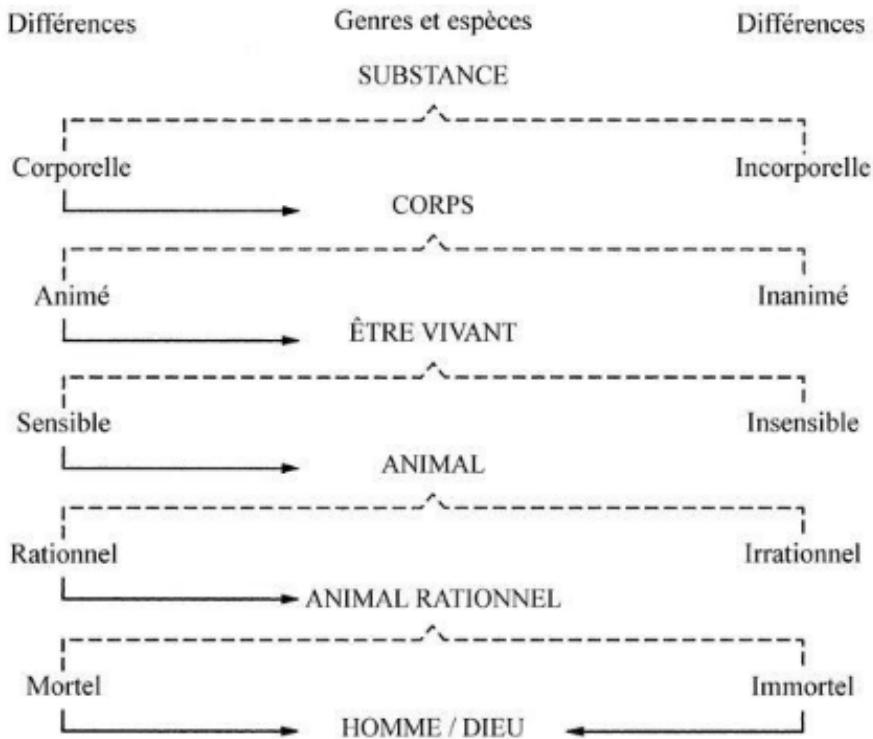


FIGURE 2.2 – Arbre porphyrien prenant en compte les différences [Source : Eco (Umberto), *De l'arbre au labyrinthe : études historiques sur le signe et l'interprétation*], trad. par Hélène Sauvage, 1 t., Paris, 2010]

doit donc être utilisé plusieurs fois dans l'arbre, ce qui le rend infini, et l'établissement d'un dictionnaire impossible à réaliser (voir Figure 2.3 : Infinitude de l'arbre de Porphyre).

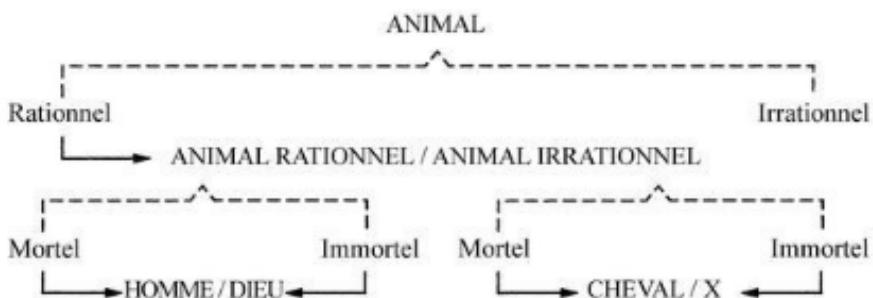


FIGURE 2.3 – Infinitude de l'arbre de Porphyre [Source : Eco (Umberto), *De l'arbre au labyrinthe : études historiques sur le signe et l'interprétation*], trad. par Hélène Sauvage, 1 t., Paris, 2010]

Face à cette impossibilité de décrire le monde avec des divisions uniques dans un seul arbre, c'est à dire d'établir un dictionnaire universel, absolu et global, la seule solution paraît être la création d'un nombre d'arbres infinis, composés de propriétés s'articulant selon le contexte et le domaine d'utilisation de l'arbre : d'un seul arbre insaisissable, une forêt réorganisable à l'envi et à l'infini est apparue, laissant le choix à l'utilisateur de

l'arbre utilisé selon le sujet.

2.1.2 L'encyclopédisme (Antiquité - Moyen-Âge) : la recherche d'un arbre global mimant le monde réel

L'utopie de saisie totale du monde se retrouve dans l'encyclopédisme, dès l'*Historia naturalis* de Pline l'ANCIEN. Sur le même principe que l'arbre porphyrien, la hiérarchie de l'index de cette encyclopédie de 37 volumes part de l'original vers le dérivé, du naturel à l'artifice : « Une encyclopédie, pour s'organiser, tente de suivre le modèle de l'arbre — qui est toujours plus ou moins consciemment celui de la subdivision binaire d'un arbre porphyrien »⁶. Cependant, l'index d'une encyclopédie se distingue des termes d'un arbre porphyrien en ce qu'il est défini dans un autre développement — un article d'encyclopédie —, alors que les termes de l'arbre de Porphyre ne peuvent pas être définis par la suite.

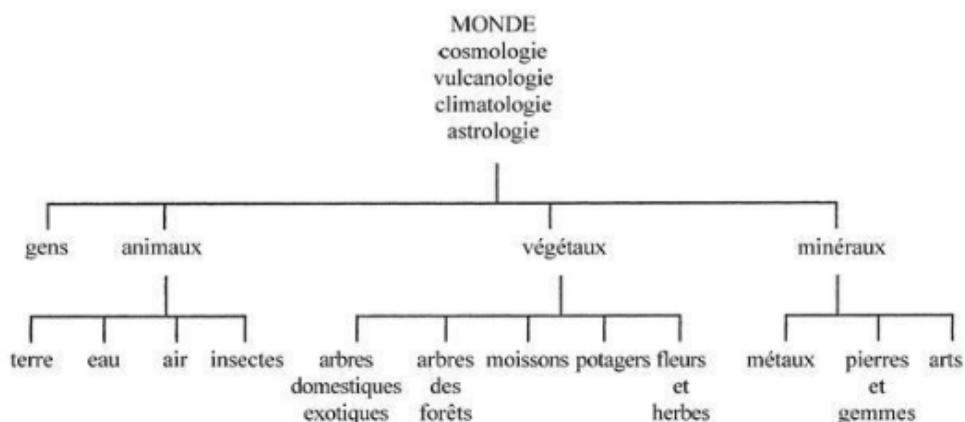


FIGURE 2.4 – Extrait de l'arborescence de l'index de Pline l'ANCIEN

Avec le passage au christianisme, l'encyclopédisme doit décrire les textes sacrés et non plus le monde. Ainsi, des éléments moralisateurs et allégoriques se retrouvent dans les index, devant les éléments matériels du monde⁷. À partir du XIII^{ème} siècle, les encyclopédies montrent l'ordre qui les dirige : cela conduit à *L'arbre de science* de Raymond LULLE qui créé seize arbres représentant l'Être, chacun représentant un savoir différent en se divisant en sept parties (racines, tronc, branches, rameaux, feuilles, fleurs, fruits)⁸. Contrairement à l'arbre de Porphyre qui est un arbre vide que l'on peut remplir selon le contexte, les arbres que propose Raymond LULLE sont pleins et ont pour vocation de décrire et de classer le monde, la Grande Chaîne de l'Être.

6. *Ibid.* Voir Figure 2.4 : Extrait de l'arborescence de l'index de Pline l'ANCIEN

7. La tradition moralisatrice encyclopédique naît avec le *Physiologos* d'un auteur grec et s'inspire de l'œuvre de Pline l'ANCIEN, et se poursuit tout au long du Moyen-Âge avec les *Étymologies* d'Isidore DE SÉVILLE notamment.

8. *Ibid.*, chap.10.

2.1.3 Influences : une diversité de référentiels hiérarchiques

La pensée aristotélicienne puis le commentaire porphyrien ont produit une tradition de hiérarchisation du monde qui s'est poursuivie pendant plus d'un millénaire, sans cesse confrontée à l'impossibilité d'une description totale de ce monde. La multiplicité des arbres est, chez Umberto ECO puis dans celle de Raymond LULLE, la conclusion de leur réflexion. L'influence de cette tradition de description est sensible jusqu'à aujourd'hui, notamment dans le domaine de l'indexation et de la bibliothéconomie.

En effet, une diversité de référentiels est apparue, chacun étant dérivé d'un arbre. Des schémas de classification sont définissables à l'infini, emboîtant les genres, les espèces et les différences⁹. La taxonomie naît de ce modèle d'arbre : la taxonomie n'a pas pour but de dire comment repérer le concept décrit, elle permet seulement de classer en renvoyant, pour chaque noeud, vers un autre chapitre où l'on décrit ces propriétés. La taxonomie, bien qu'historiquement appliquée aux sciences de la terre, a été reprise par Melvil DEWEY dans sa classification décimale en 1876.

Définie comme un « classement hiérarchique de termes préférentiels » par Louis ROSENFIELD et Peter MORVILLE¹⁰, la taxonomie ne veut pas définir, mais simplement permettre l'utilisation correcte et logique du terme par l'attribution de catégories et l'utilisation exclusive de relations hiérarchiques.

Les *thesauri*¹¹ utilisent plus de relations et de types de termes, de manière à indexer des contenus avec des mots-clés et à faciliter la recherche. Ce vocabulaire contrôlé hiérarchique reste proche du langage naturel en y intégrant les variantes, les synonymes, les descriptions, les traductions et les équivalences.

Pour avoir une plus grande formalisation du thésaurus, il faut utiliser une ontologie. Cette ontologie est la spécification formelle d'un espace de noms, d'un domaine particulier de la connaissance¹². Elle identifie alors les objets à décrire, leurs relations au

9. « Un simple artifice classificatoire consiste à emboîter des genres, des espèces et des différences sans en expliquer le *definiendum* » in *Ibid.*, chap.1

10. L. Rosenfeld, P. Morville et J. Arango, *Information architecture for the World Wide Web...*

11. Ils sont décrits comme une « liste organisée de termes contrôlées et normalisés (descripteurs et non-descripteurs) servant à l'indexation des documents et des questions dans un système documentaire » dans Danièle Dégez et Dominique Menillet, *Thésauroglossaire des langages documentaires : un outil de contrôle sémantique*, Paris, 2001 (Collection Sciences de l'information), URL : <https://catalogue.bnf.fr/ark:/12148/cb37703277d>. Peu formels, ils sont néanmoins le vocabulaire le plus utilisé pour l'indexation. L'un des *thesauri* les plus utilisés est le General Multilingual Environmental Thesaurus (GEMET)(*General Multilingual Environmental Thesaurus*, URL : <https://www.eionet.europa.eu/gemet/en/groups/> (visité le 03/09/2020)). Le est disponible en plus de trente langues et diffusé par l'Agence européenne de l'Énergie. Voir section 2.2 : Le *thésaurus*, vocabulaire contrôlé hiérarchique le plus fréquent.

12. L'une des ontologies les plus utilisées, notamment dans le web sémantique, est Friend of a Friend

sein de ce domaine ainsi que leurs propriétés. L'ontologie n'est pas utilisée directement dans l'indexation ou la recherche, elle est d'abord utilisée pour instancier et raisonner, en s'éloignant du langage naturel avec l'utilisation d'identifiants techniques.

Les taxonomies, les *thesauri* ainsi que les ontologies héritent tous du modèle de l'arbre, la description ou la classification par la hiérarchie étant la plus efficace pour ces besoins. Ces vocabulaires sont les plus complexes par les relations qui les composent. Louis ROSENFIELD et Peter MORVILLE¹³ considèrent l'anneau de synonymie comme le plus simple des vocabulaires, avec des relations d'équivalence, alors que les fichiers d'autorité et les taxonomies, fonctionnant sur la hiérarchie, sont plus complexes. Les *thesauri* et les ontologies sont plus complexes encore puisqu'ils sont constitués de relations hiérarchiques et associatives.

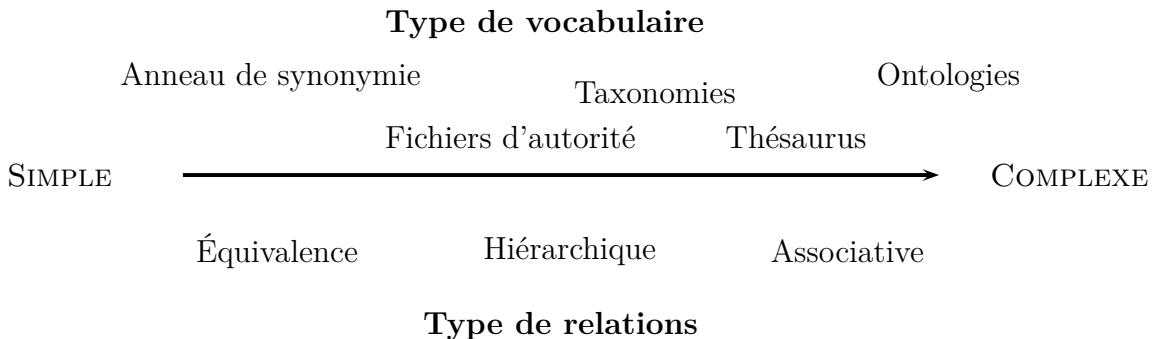


FIGURE 2.5 – Classification des vocabulaires selon leur complexité [d'après ROSENFIELD (Louis), MORVILLE (Peter) et ARANGO (Jorge), *Information architecture for the World Wide Web*, OCLC : 922954742, Beijing, 2015]

2.2 Le *thesaurus*, vocabulaire contrôlé hiérarchique le plus fréquent

Né dans les années 1950 aux États-Unis, le thésaurus n'a été adopté massivement qu'avec l'apparition de l'informatique. C'est un langage combinatoire, une liste organisée de termes normalisés et contrôlés, qui permet de faire le lien entre le langage naturel de l'homme et le nécessaire besoin d'avoir un langage contrôlé pour les ressources. La sélection d'un terme lors de l'indexation permet de sélectionner un concept lui-même

(FOAF). permet la description précise des personnes. Voir *FOAF Vocabulary Specification*, URL : <http://xmlns.com/foaf/spec/> (visité le 03/09/2020)

13. L. Rosenfeld, P. Morville et J. Arango, *Information architecture for the World Wide Web...* Voir Figure 2.5 : Classification des vocabulaires selon leur complexité

décrit par plusieurs termes (synonymes, équivalents, traductions). Ainsi, les institutions patrimoniales se sont emparées de cet outil, adaptable au domaine de chacune : l'INA possède un thésaurus orienté vers l'audiovisuel, la Cinémathèque française un thésaurus orienté vers le cinéma.

2.2.1 Types de structure

Le type de thésaurus le plus utilisé est celui constitué d'une hiérarchie simple¹⁴. L'INA possède un thésaurus de noms communs formé sur cette hiérarchie simple à unique ascendance¹⁵, c'est à dire qu'un terme est nécessairement descendant d'une seule classe, il ne peut pas hériter de deux caractéristiques différentes, ce qui le rapproche de la taxonomie¹⁶.

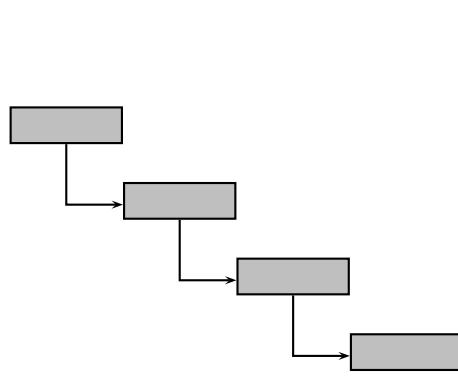


FIGURE 2.6 – Le modèle taxonomique

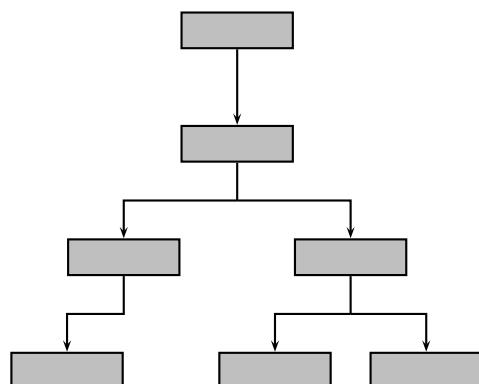


FIGURE 2.7 – Le modèle du thésaurus simple

Comparaison entre le modèle taxonomique et celui du thésaurus à hiérarchie simple [d'après ROSENFELD (Louis), MORVILLE (Peter) et ARANGO (Jorge), *Information architecture for the World Wide Web*, OCLC : 922954742, Beijing, 2015]

De manière à exprimer la descendance depuis plusieurs caractéristiques, un thésaurus polyhiérarchique existe. Il permet de définir et d'accepter plus de termes contrôlés que le thésaurus simple. En effet, par la combinaison des termes ascendants, un même terme peut avoir deux ascendance différentes. Peter MORVILLE et Louis ROSENFELD prennent un exemple médical pour illustrer ce type particulier de thésaurus.

Enfin, comme nous l'avons évoqué précédemment¹⁷, le seul arbre possible est un arbre multiple, adapté à son contexte. Ainsi, des *thesauri* à facettes existent, reflétant les multiples dimensions thématiques que peuvent contenir les documents ou les éléments : un terme se retrouve alors dans plusieurs arbres, multipliant les points d'accès. Plusieurs

14. La typologie des *thesauri* décrite par la suite est présente chez *Ibid.*

15. Voir Annexe E : Le thésaurus de noms communs de l'INA

16. Voir Figure 2.6 : Le modèle taxonomique et Figure 2.7 : Le modèle du thésaurus simple.

17. Voir section 2.1 : L'arbre de Porphyre : origines et influences.

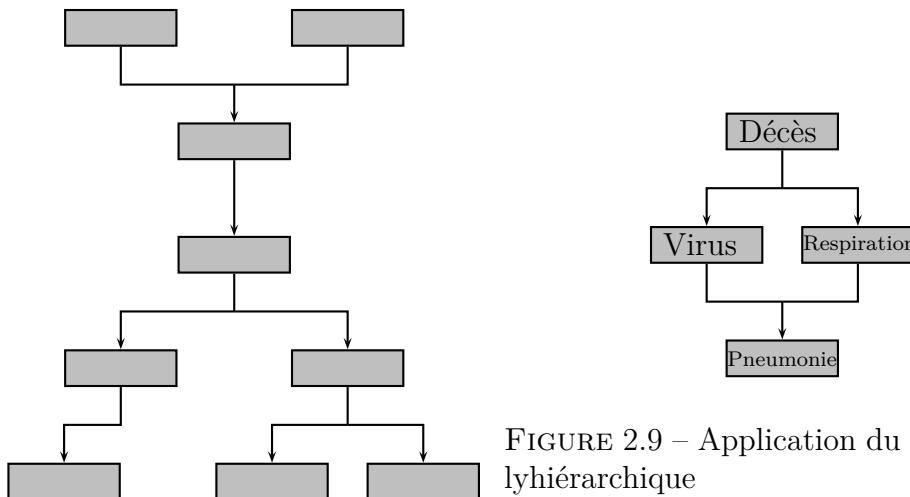


FIGURE 2.8 – Le modèle polyhiérarchique

Le modèle du thésaurus polyhiérarchique [d’après ROSENFELD (Louis), MORVILLE (Peter) et ARANGO (Jorge), *Information architecture for the World Wide Web*, OCLC : 922954742, Beijing, 2015]

thesauri simples sont par conséquent créés, permettant la description de l’ensemble de ces dimensions.

2.2.2 Relations entre les termes

La force du thésaurus ne réside pas seulement dans l’enchaînement d’ascendances et de descendances. Les relations établies entre les termes sont essentielles pour permettre le lien entre le langage humain naturel et le besoin de contrôle imposé par l’indexation et la recherche : un thésaurus est « un vocabulaire contrôlé dans lequel les relations d’équivalence, de hiérarchie et d’association sont correctement identifiées de manière à permettre une meilleure récupération »¹⁸.

Les relations créées précisent le sens de chaque vedette par comparaison aux vedettes de sens voisin, elles permettent de naviguer entre ces vedettes pour affiner sa recherche, l’élargir ou bien la réorienter. La hiérarchisation et l’établissement de liens permet de passer à une navigation sémantique, alors que les simples vocabulaires contrôlés évoqués au Chapitre 1 : Le référentiel comme clé ne permettaient qu’une navigation par mots.

La première relation est la relation d’équivalence. Elle connecte le terme préférentiel — le terme principal de la vedette — avec ses variantes : les synonymes, les acronymes, les abréviations, les variantes lexicales ou les différences de graphie sont ainsi incorporés

18. *Ibid.*

au thésaurus comme variantes. Cette relation¹⁹ est une relation horizontale, d'égalité, comme dans l'anneau de synonymie. Dans l'Annexe E : Le thésaurus de noms communs de l'INA, le terme « Cadreur », qui est le terme préférentiel, a deux variantes — ou termes « Employés pour », « Cameraman » et « Opérateur de prise de vue ».

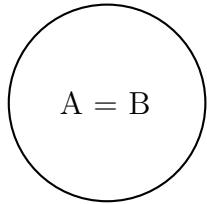


FIGURE 2.10 – Relation d'équivalence

Le second type de relation est la relation associative. Comme la relation d'équivalence, elle est horizontale. Elle permet d'exprimer la proximité sémantique entre deux termes : dans RAMEAU, la vedette « Télévision » possède quarante relations d'association avec d'autres vedettes, comme « Industrie de la télévision ». L'association²⁰ n'est pas une relation de stricte égalité, elle indique le partage sémantique d'une partie de leur définition. Cette relation permet l'élargissement d'une recherche depuis une vedette.

Le dernier type de relation est hiérarchique. Il est le plus utilisé car il permet l'expression de nombreuses relations du langage naturel :

- la relation génétique — la plus fréquente — peut ainsi être exprimée. Le sens du terme générique est inclus dans celui du terme spécifique : la vedette RAMEAU « Radiodiffusion » est l'un des termes génériques de « Télévision » qui est elle-même terme générique de « Chaînes de télévision » notamment. Chacune de ces vedettes est décrite par son ascendance et sa descendance.
- la relation d'appartenance — ou de regroupement — est possible ;
- de même que la relation partitive

La définition de cette relation hiérarchique²¹ permet l'expression de caractéristiques et de relations du langage naturel infinies. La recherche d'une vedette peut alors être affinée — quand l'utilisateur passe d'une vedette générique à une vedette spécifique — ou bien élargie — quand il passe d'une vedette spécifique à une vedette générique.

Alors, chaque terme devient le centre de son propre réseau et construit un nouvel arbre, entièrement né de son contexte.

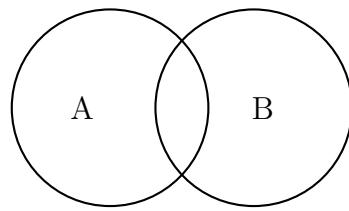


FIGURE 2.11 – Relation d'association

19. Voir Figure 2.10 : Relation d'équivalence

20. Voir Figure 2.11 : Relation d'association.

21. Voir Figure 2.12 : Relation de hiérarchie.

2.2.3 Utiliser la précoordination pour les relations complexes

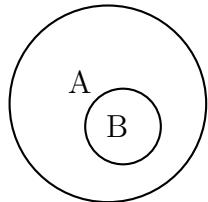


FIGURE 2.12 – Relation de hiérarchie

L'inconvénient du thésaurus comme évoqué précédemment est l'impossibilité pour l'utilisateur de feuilleter l'index : « Télévision » et « Chaînes de télévision », bien qu'étant proches, ne seraient pas au même endroit dans l'index. Pour faciliter la navigation de l'utilisateur, les mots-clés sont coordonnés avant l'utilisation par l'utilisateur pour former une vedette-matière construite (comme dans le cas de RAMEAU) : une vedette principale constitue la tête de la vedette, puis des subdivisions la complètent²². Une vision globale est ainsi offerte et permet une précision du sujet des facettes ainsi qu'une limitation du bruit : Plantes – Parasites – Plantes-hôtes » est ainsi séparée de « Plantes parasites ».

Les différentes structures de *thesauri* et leurs multiples relations permettent un modèle de classification, de combinaison et de description des termes efficace, à la fois proche du langage naturel mais en s'en éloignant par le formalisme et le contrôle des termes. Chaque vedette est le centre de son propre référentiel, dirigeant vers des variantes, des vedettes proches ou en relation.

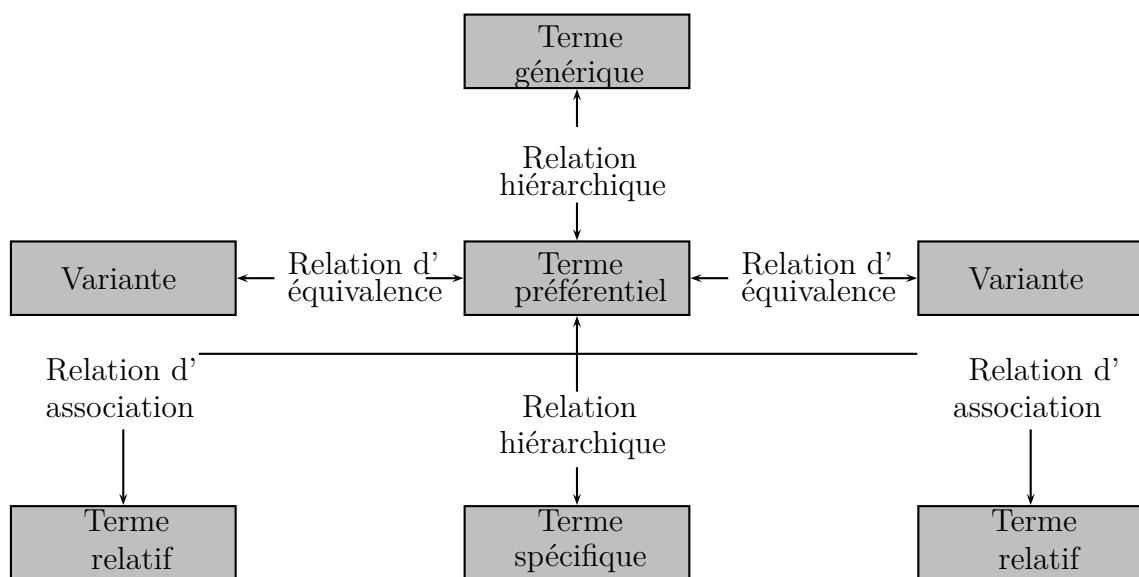


FIGURE 2.13 – Modélisation d'une vedette de thésaurus [d'après ROSENFELD (Louis), MORVILLE (Peter) et ARANGO (Jorge), *Information architecture for the World Wide Web*, OCLC : 922954742, Beijing, 2015]

22. Dans « Plantes – Parasites – Plantes-hôtes », « Plantes » est la tête de vedette, complétée par deux subdivisions.

2.3 Passer du texte libre à un vocabulaire contrôlé : aligner des notes qualité et un thésaurus de noms communs

Dans la description de documents audiovisuels — comme dans celle d'autres documents patrimoniaux —, désigner des personnes est indispensable. Pour enrichir le seul état civil de la personne, plusieurs moyens peuvent être utilisés :

- rédiger un texte libre décrivant les caractéristiques de la personne, ses fonctions, ses dates de naissance et de mort, Cette solution pose la problématique de la structuration des données : un texte libre n'est pas lisible par une machine ; son accès est par conséquent restreint.
- utiliser un vocabulaire contrôlé et sélectionner les termes correspondant à la personne. Cependant, en fonction du niveau de précision souhaité, ce vocabulaire doit être plus ou moins précis, rendant, dans le cas d'une grande précision, la description longue et fastidieuse.
- définir des champs essentiels à la description de cette personne, et rédiger un texte libre pour les informations supplémentaires. De même que dans le premier cas, le texte libre appauvrit l'effort de structuration de la description.

Face à ces difficultés, les documentalistes de la DDCOL à l'INA créent des vedettes de personnes selon une succession de champs (sexe, date de naissance, date de mort, ...) et de notes, dont une note qualité qui est régie par un guide de rédaction. Cette note qualité a pour but de décrire en quelques mots les fonctions de la personne et le lieu d'exercice. Contrairement au nom et au prénom de la personne, la note qualité n'est pas une clé dans les données : n'étant pas un point d'accès, elle peut être structurée et rédigée en texte libre.

Dans le cadre de la migration des données de la DDCOL dans le *Lac de données*, un alignement de ces notes qualité est nécessaire avec le thésaurus des noms communs qui existe parallèlement, notamment pour enrichir le thésaurus des fonctions des notes qualité qui n'y existent pas.

2.3.1 Contrôler du texte libre

La note qualité est rédigée selon des règles définies au préalable par les documentalistes. Cependant, la rédaction en texte libre conduit à l'apparition d'erreurs humaines, comme les erreurs de graphie, de grammaire ou de ponctuation. En effet, une note qualité peut avoir deux formes :

- Fonction1, fonction2, Pays
- Homonymes : 1 - Fonction1, fonction2, Pays ; 2 - Fonction1, fonction2,
Pays

Ainsi, l'oubli d'une ponctuation, ou son inversion, conduit à rendre la note qualité non conforme aux règles et, par conséquent, à rendre son traitement plus difficile voire impossible. De plus, les différences de graphie liées au masculin et au féminin, ainsi qu'au singulier et au pluriel, rendent ces notes qualité très différentes.

De manière à pouvoir les aligner avec le thésaurus des noms communs, un premier traitement est nécessaire, pour extraire et normaliser les fonctions. Le logiciel ETL (*Extract Transform Load*)²³ Talend Big Data Platform permet ce premier traitement.

La première étape consiste à scinder chaque note selon les fonctions et les pays : le point sépare ces deux éléments et permet cette scission. Ainsi, la fonction extraite de « Historien, musicologue. France » est « Historien, musicologue » alors que la note qualité « Journaliste, France » ne peut pas être scindée. Une seconde scission intervient par la suite de manière à récupérer chaque fonction une à une, passant de « Historien, musicologue » à « Historien » et « musicologue ».

Quand les fonctions sont récupérées, le contrôle des termes peut avoir lieu selon plusieurs choix à effectuer en amont :

- le choix du genre doit être effectué pour éviter les termes équivalents dans le sens mais différents en graphie
- le choix du nombre
- la gestion de la ponctuation propre aux fonctions comme les traits d'union
- la gestion de l'accentuation

Pour normaliser le plus possible, le choix du masculin singulier, de la suppression de toute la ponctuation et de l'accentuation a été effectué. Pour le choix du genre, le nombre des exceptions comme « musée », portant une terminaison du féminin, étant plus rares que le nombre de tous les féminins, le choix du masculin s'est imposé pour normaliser le maximum de fonctions. La Figure 2.14 montre une dernière normalisation à effectuer : la suppression des *stopwords*, effectuée en Figure 2.15.

Après la normalisation, les fonctions sont suffisamment contrôlées et proches des règles d'un thésaurus pour être alignées. Cependant, nous pouvons observer que les erreurs humaines de graphie, comme l'oubli d'un « s » dans « dessinateur », restent et ne pourront par conséquent pas être alignées. Le traitement correct de l'ensemble des notes en texte libre reste impossible à cause des erreurs introduites par l'homme.

Enfin, les notes qualité de l'INA comprennent également des qualités ne décrivant pas directement la personne, mais définissant cette personne par un lien avec un fait. C'est

23. Un ETL permet de migrer des données depuis une source vers une cible, en leur appliquant des traitements avant de les charger dans la cible.

	Note qualité	Fonction normalisée
1	Dessinateur de presse. France	dessinateur de presse
2	Dessinateur, scénariste. France	dessinateur
2	Dessinateur, scénariste. France	scénariste
3	Desinateur, illustrateur, graphiste. France	desinateur
3	Desinateur, illustrateur, graphiste. France	illustrateur
3	Desinateur, illustrateur, graphiste. France	graphiste
4	Dessinatrice, animatrice. France	dessinatrice
4	Dessinatrice, animatrice. France	animateur

FIGURE 2.14 – Données d'exemple de notes qualité avec la fonction de Réalisateur

	Note qualité	Fonction normalisée
1	Dessinateur de presse. France	dessinateur presse
2	Dessinateur, scénariste. France	dessinateur
2	Dessinateur, scénariste. France	scénariste
3	Desinateur, illustrateur, graphiste. France	desinateur
3	Desinateur, illustrateur, graphiste. France	illustrateur
3	Desinateur, illustrateur, graphiste. France	graphiste
4	Dessinatrice, animatrice. France	dessinatrice
4	Dessinatrice, animatrice. France	animateur

FIGURE 2.15 – Données d'exemple de notes qualité avec la fonction de Réalisateur, après normalisation des fonctions

le cas des faits divers, des attentats, des affaires judiciaires dans lesquels une personne peut être impliquées comme victime, accusé, témoin, ... ; c'est le cas également des indications de filiation et de généalogie avec lesquelles une personne est seulement désignée, sans apporter de précisions sur ses véritables fonctions²⁴. Ces parties de notes qualité — ou bien la totalité de ces notes — ne décrivant pas la fonction de la personne et n'allant pas trouver d'équivalent dans le thésaurus, elles sont écartées du traitement.

	Note qualité	Fonction normalisée
1	Affaire transformateur électrique à Clichy-sous-Bois, victime. France	
2	Attentat, Paris novembre 2015, suspecte. France	
3	Ecrivain, fils de Victor Hugo. France	ecrivain
3	Ecrivain, fils de Victor Hugo. France	

FIGURE 2.16 – Données d'exemple de notes qualité sans fonctions

24. Voir Figure 2.16 : Données d'exemple de notes qualité sans fonctions.

2.3.2 Aligner les extractions en langage naturel avec un thésaurus de noms communs

Avec le premier traitement de normalisation des fonctions, les notes qualité sont sorties du langage naturel de manière à pouvoir être contrôlées dans un vocabulaire plus strict. L'alignement avec le thésaurus de noms communs peut alors être réalisé²⁵. Ce thésaurus est classé dans un ordre hiérarchique, mais l'accès par des termes ascendants est difficile pour l'alignement : les termes génériques sont souvent des noms qui ne sont pas des fonctions, ce qui rend leur alignement impossible. Ainsi, le terme « Dessinateur » a pour ascendance « \$art et culture/arts plastiques/dessin » : « Dessin » ou « Arts plastiques » ne sont pas des fonctions. L'ensemble des alignements est par conséquent réalisé avec les termes préférentiels les plus bas dans l'arborescence. Le thésaurus contenant également des synonymes²⁶, ces derniers sont utilisés dans l'alignement de manière à réduire encore l'impact du langage naturel des notes qualité sur la qualité de l'alignement.

Cette phase d'alignement est également réalisée avec Talend grâce à une successions de jointures²⁷. Les fonctions strictement égales au terme préférentiel du thésaurus sont ainsi alignées, ainsi que celles qui commencent par un terme du thésaurus. Cette étape de l'alignement montre les difficultés posées par l'utilisation du texte libre dans la description et la gestion impossible des coquilles, bien que parfois très proche du terme exact²⁸.

Face à ces difficultés et au nombre peu élevé des alignements qui résultent de cette étape, l'utilisation des synonymes peut apporter des résultats supplémentaires : l'entrée « Cuisinier » du thésaurus de noms communs comprend un synonyme, « Chef de cuisine ». Cependant, le nombre des synonymes est réduit, et des alignements sont ici aussi non réalisés²⁹.

Enfin, le cas de « Chef cuisinier » montre la nécessité d'utiliser le second terme de l'expression de la fonction³⁰ : cette dernière étape de l'alignement permet l'alignement des fonctions commençant par des termes polysémiques comme « Chef », « Directeur », « Maître »,...

25. De manière à avoir la même normalisation de chaque côté de l'alignement, le thésaurus a subi le même traitement que les notes qualité, avec l'application des mêmes règles.

26. Voir les termes Employés pour dans l'Annexe E : Le thésaurus de noms communs de l'INA.

27. Ici, les jointures sont des *inner join* pour aligner sur la similarité entre les deux côtés — fonctions issues de la note qualité, et termes du thésaurus — , ou bien des comparaison effectuées à partir du début de la fonction issue des notes qualité — « Illustrateur de presse » pourra ainsi correspondre au terme « Illustrateur » du thésaurus.

28. Voir l'exemple de l'alignement du terme « Journaliste » Annexe F : Aligner les fonctions de « Journaliste » des notes qualité avec le thésaurus des noms communs de l'INA.

29. Voir Figure 2.17 : Utilisation des synonymes pour l'alignement du terme « Cuisinier »

30. Voir Figure 2.18 : Gestion de la polysémie dans l'alignement du terme « Cuisinier »

	Note qualité	Fonction normalisée	Terme du thésaurus	Vedette du thésaurus
1	Chef cuisinier. France	chef cuisinier	cuisinier	1
2	Cuisinière. France	cuisinier		
3	Cheffe cuisinière. France	chef cuisinier		
4	Cheffe cuisinier. Italie	chef cuisinier		
5	Chef de cuisine. France	chef cuisine	chef de cuisine	1

FIGURE 2.17 – Utilisation des synonymes pour l’alignement du terme « Cuisinier »

	Note qualité	Fonction normalisée	Terme du thésaurus	Vedette du thésaurus
1	Chef cuisinier. France	chef cuisinier	cuisinier	1
2	Cuisinière. France	cuisinier	cuisinier	1
3	Cheffe cuisinière. France	chef cuisinier	cuisinier	1
4	Cheffe cuisinier. Italie	chef cuisinier	cuisinier	1
5	Chef de cuisine. France	chef cuisine	chef de cuisine	1

FIGURE 2.18 – Gestion de la polysémie dans l’alignement du terme « Cuisinier »

2.3.3 Classer selon le thésaurus

L’utilisation des relations d’association a permis d’aligner les fonctions avec les termes du thésaurus. Les relations de hiérarchie avec les termes génériques permettent de classer ces fonctions. Ainsi, elles sont utilisées pour définir huit catégories de rattachement dans les fonctions extraites des notes qualité, de manière à les classer selon le thésaurus. Dans le thésaurus des noms communs, huit termes permettent de rattacher l’ensemble des termes spécifiques, souvent avec des niveaux de hiérarchie intermédiaires³¹. Ces termes de catégorisation sont des facettes : ils ne sont pas attribuables directement à un concept à indexer, ils permettent le seul classement.

Cette opération de classement des fonctions des notes qualité selon l’arborescence du thésaurus permet, au-delà de l’ajout sémantique sur les termes alignés, de repérer les termes qui n’ont pas été alignés et d’en comprendre les raisons :

- Des noms communs ne correspondant pas à des fonctions sont présents dans les notes qualité. Ainsi, des termes comme « cirque pinder » ou « clip » ne trouveront pas d’équivalence dans le thésaurus.
- des noms trop spécifiques ne sont également pas présents : « chemisier » est une fonction spécifique que les documentalistes pourront créer si nécessaire grâce à ce repérage dans les notes qualité.
- Des erreurs introduites par accident par l’homme empêche certains alignements : c’est le cas par exemple de « chercher » qui a un équivalent « chercheur » dans le thésaurus. Il est difficile de repérer et de corriger ces erreurs automatiquement.

31. Ces huit catégories sont : « Art et culture », « communication diffusion traitement information », « sciences », « sciences humaines », « sport », « vie économique », « vie quotidienne habitat alimentation et loisirs » et « vie sociale ».

- La présence d'une documentation d'aide au catalogage et à l'indexation permet d'introduire de nouvelles règles dans la classification automatique : ainsi, « designer interieur » peut être classifié dans la facette « Art et culture » car la documentation l'indique ; cependant, le terme « designer » étant absent du thésaurus, il ne peut pas être aligné.

L'utilisation d'un thésaurus permet d'aligner des termes ensemble et de relier du texte libre avec un vocabulaire contrôlé de manière à disposer d'un vocabulaire commun de description. Plus encore, la hiérarchie d'un thésaurus permet la classification d'un ensemble de concepts — ici les fonctions — selon quelques catégories globales. Le thésaurus a par conséquent la double fonction d'offrir un enrichissement du terme préférentiel par ses relations d'association, et de proposer une classification par ses relations hiérarchiques.

Aligner du texte libre avec un thésaurus nécessite plusieurs étapes et la prise en compte des différences de langage — l'un étant un contrôle minimal du langage humain naturel, l'autre un vocabulaire contrôlé natif — :

- une normalisation est d'abord nécessaire de chaque côté de l'alignement selon les mêmes règles ;
- puis un alignement selon l'exactitude avec le terme préférentiel peut être réalisé
- suivi par un autre alignement selon l'exactitude avec une variante du terme préférentiel ;
- ensuite, aligner selon l'exactitude du commencement de la fonction avec le terme préférentiel est possible,
- ainsi qu'aligner selon l'exactitude du commencement de la fonction avec une variante du terme préférentiel
- pour enfin aligner selon l'exactitude du deuxième terme non polysémique de la fonction avec le terme préférentiel

Chapitre 3

Les référentiels à l'INA

« Conserver, c'est d'abord faire en sorte que chaque minute audiovisuelle qui passe soit une archive. Si pour le téléspectateur la diffusion d'une émission renvoie au moment présent, pour l'INA, il s'agit déjà d'une parcelle de mémoire à conserver¹. »

Historiquement créé avec la scission de l'ORTF en 1974 pour la collecte des archives audiovisuelles, la recherche et la formation professionnelle, l'INA a subi un bouleversement dans les années 1990 avec la loi du 20 juin 1992 sur le dépôt légal audiovisuel instaurant le dépôt légal des radios publiques — en 1994 —, des télévisions nationales hertziennes — en 1995 —, et de la télévision par câble et numérique et de radios privées — en 2002, captées directement, permettant alors de s'affranchir du dépôt physique par les producteurs et d'augmenter en masse les données de l'Institut.

Défini comme un ÉPIC dans la loi de création de 1974, l'INA a, en plus des missions patrimoniales décrites ci-dessus, des missions commerciales comme la commercialisation des archives ou bien la vente de services auprès de producteurs audiovisuels — l'INA devient alors un tiers-archiveur par le biais de mandats signés avec ces producteurs pour la conservation et/ou la commercialisation de leurs archives. Des compétences juridiques sont ainsi indispensables à l'INA pour cette commercialisation et le reversement des droits aux ouvrants- et ayants-droits.

Chacune de ces missions a des besoins différents et dirige la structure des données ainsi que leur usages. La création et l'usage de référentiels sont alors différents selon la mission du département gestionnaire des données : la DDCOL est en charge de la mission patrimoniale, tandis que la DJ est en charge des aspects commerciaux et juridiques.

1. Emmanuel Hoog, *L'INA*, ISSN : 0768-0066, 2006 (Que sais-je?), URL : <https://www-cairn-info.proxy.chartes.ps1.eu/l-ina--9782130545453.htm> (visité le 05/07/2020), p.45.

3.1 De multiples fonds à décrire

3.1.1 Les archives professionnelles

Après l'éclatement de l'ORTF, des fonds divers deviennent la propriété de l'INA et participent à la diversité des fonds audiovisuels conservés à l'INA. Ainsi, les « Actualités françaises » — la presse filmée diffusée dans les selles de cinéma — ont été transférées à l'INA en 1974. Les grands moments des débuts de la télévision — le premier journal télévisé, les premières grandes émissions ou les magazines de reportage — sont conservés et participent, avant l'instauration du dépôt légal, à la mémoire de l'audiovisuel français ; de même pour les fonds radio qui retracent les grands moments historiques depuis les années 1940, et qui deviennent de plus en plus complets avec la généralisation de la bande magnétique à partir du milieu des années 1950.

Les archives professionnelles de l'INA ne sont que des archives : elles ne couvrent pas l'ensemble de la production audiovisuelle depuis les années 1940 ou la création de l'Institut. Des tranches horaires de diffusion télé- ou radiodiffusées ne sont ainsi pas conservées et peu de traces demeurent alors de la production audiovisuelle, notamment avant l'arrivée du kinéscope. Pour combler ces manques, l'INA dispose d'un fonds photographique créé à partir des services de l'ORTF ou de l'INA et portant sur la réalisation des émissions et des tournages.

Enfin, des délégations régionales se chargent de la conservation et de la communication des archives télévisées et radiodiffusées des stations régionales — apparues dans les années 1950 : la vie et l'histoire des régions sont ainsi couvertes par l'INA.

Les fonds d'archives professionnelles de l'INA sont conséquents et divers, témoins de la vie et de la société française depuis l'Après-Guerre. Irremplaçables, leur description n'en est pas moins difficile par la diversité des sujets évoqués ou présents dans les documents.

3.1.2 Les fonds issus du dépôt légal

Depuis la loi sur le dépôt légal de l'audiovisuel de 1992, l'INA en est le dépositaire. À partir de 1995, l'INA enregistre la globalité de la programmation des stations de Radio-France — France-Inter, France-Musique, France-Culture, France-Info et France-Bleue —, enregistrement étendu en 2001 aux stations privées généralistes comme RTL ou NRJ².

2. En 2020, l'ensemble des stations captées au titre du dépôt légal par l'INA est décrit dans l'Annexe G : Les captations directes réalisées par l'INA au titre du dépôt légal (Figure G.1 : Stations de radio captées au titre du dépôt légal)

Pour les programmes télévisés, le dépôt légal ne concerne d'abord — entre 1995 et 2001 — que les sept chaînes principales — TF1, France 2, France 3, Canal +, M6, Arte, France 5 — et leurs programmes en première diffusion. La captation directe et intégrale des chaînes n'apparaît qu'en 2002 et est élargie à douze autres chaînes. Enfin, depuis 2005, les chaînes de la Télévision numérique terrestre (TNT) sont toutes captées³.

La diversité des fonds d'archives, la captation directe en intégralité des chaînes de télévision et de radio, ainsi que la captation de sites web, plateformes ou comptes de réseaux sociaux au titre du dépôt légal audiovisuel, représentent une masse très importante de documents à conserver et de données. En 2019⁴, l'INA conserve 20 873 143 heures de programmes de télévision et de radio, dont plus de 18 millions captés par le dépôt légal. 1,2 million de photos s'ajoutent à ces documents. La majorité de ces documents, issus du dépôt légal, sont destinés à une gestion patrimoniale et à une valorisation dans l'INATHÈQUE, alors que les documents des archives professionnelles sont destinées à la valorisation commerciale au travers notamment le site INAMediaPro destinés aux professionnels.

3.2 Un système documentaire pluriel répondant aux besoins

La masse des documents, l'évolution de leur récupération auprès des producteurs et leurs usages divers conduisent à la création d'un système documentaire pluriel, créé à partir des besoins et non des données. Les deux usages commerciaux et patrimoniaux des documents⁵, évoqués précédemment, dirigent le nombre des bases de données, leur structure et le partage de référentiels. Avant le projet du *Lac de données* lancé en 2015, le système documentaire de l'INA est pluriel, constitué de plusieurs bases de données distinctes ainsi que de plusieurs référentiels non communs.

Deux types de données sont présentes dans les bases de l'INA. D'abord, il y a du texte libre, décrivant les titres propres des documents, les titres de collections ou indi-

3. Les chaînes de télévision captées en 2002 pour le dépôt légal sont décrites dans l'Annexe G : Les captations directes réalisées par l'INA au titre du dépôt légal (Figure G.2 : Chaînes de télévision captées au titre du dépôt légal)

4. Institut national de l'Audiovisuel, *Rapport d'activités 2019*, Bry-sur-Marne, Institut National de l'Audiovisuel, 2019, p.5.

5. É. Alquier décrit ces deux usages dans un article de 2017 évoque ces usages et la plateforme qui les met en œuvre. En 2020, le service de vidéo à la demande Madelen vient s'ajouter à l'offre « grand public » de l'INA. Voir Eléonore Alquier, Jean Carrive et Steffen Lalande, “Production documentaire et usages”, *Document numérique*, Vol. 20–2 (2017), Publisher : Lavoisier, p. 115-136, URL : <https://www-cairn-info.proxy.psl.eu/revue-document-numerique-2017-2-page-115.htm> (visité le 05/07/2020).

quant un identifiant, ou bien des notes diverses ou des chiffres. Ensuite, il y a les données contrôlées, issues de référentiels et de lexiques, permettant de décrire les contenus, les particularités de ces contenus et les événements associés à ces contenus (diffusion, archivage, exploitation) Annexe C : Les types de données présents dans les bases de données de l'INA et leur rôle.

3.2.1 Les bases de données du dépôt légal (DL)

L'INA capte en permanence et en direct plus de 170 chaînes de télévision et stations de radio⁶. Ce flux ininterrompu est décrit lors du catalogage par des techniciens spécialisés dans la gestion de collections multimédia : le titre, le générique, les dates et heures de diffusion notamment sont ainsi indiqués pour chaque document, ainsi que des descripteurs pour indexer la chaîne de diffusion, les thématiques présentes, ...

Quand le document est décrit, les données complétées par le technicien de gestion des collections multimédia dans son interface graphique sont dirigées vers les bases de données du dépôt légal, scindées en quatre pour correspondre à la provenance du document. Ainsi, bien que les documents proviennent de la même source — la captation pour le dépôt légal —, ils sont éclatés dans quatre bases de données différentes pour correspondre à leur provenance :

- la base DLRADIO (Dépôt Légal de la Radio) comprend les documents diffusés en radio, sans autre distinction de provenance
- la base DLTV (Dépôt Légal de la Télévision (Nationale)) ne comprend pas l'ensemble des documents diffusés à la télévision, mais seulement les chaînes nationales
- la base DLREG (Dépôt Légal de la Télévision Régionale) comprend les documents télévisuels diffusés sur une chaîne de télévision régionale comme France3
- enfin, la base DLSAT (Dépôt Légal de la Télévision Satellite) comprend les documents diffusés sur les chaînes de télévision satellite

Cependant, malgré cette scission des données dans plusieurs bases de données, ces quatre bases partagent un même schéma pour les référentiels. Ce schéma permet de trouver des tables comprenant la signification d'identifiants de provenance de chaînes (le lien entre le code « FR5 » présent dans les données peut ainsi être établi avec son terme développé), de provenance de données, ... Ce schéma est un fournisseur de mots-clés destinés à permettre la description, l'indexation et la recherche des documents.

6. I. national de l'Audiovisuel, *Rapport d'activités 2019...*, p.5.

3.2.2 Les bases de données des archives professionnelles (DA)

Le dépôt légal se concentre sur la diffusion des documents et conserve alors l'ensemble de ce qui est diffusé à la télévision ou à la radio — les émissions, les films, les publicités, les journaux télévisés, ... — à chaque instant. Cette conservation des premières diffusions et des rediffusions permet, ainsi que l'exige le dépôt légal, d'avoir un panorama complet du paysage audiovisuel français, comme c'est le cas à la Bibliothèque nationale de France pour les imprimés ou les périodiques.

Les archives professionnelles ne sont pas soumises à cette exhaustivité : lorsqu'un producteur de contenu audiovisuel mandate l'INA pour la conservation et/ou la commercialisation de ses contenus, les données de ces contenus sont récupérées dans les bases du dépôt légal puis copiées dans celles des archives professionnelles. Ainsi, la même donnée est dupliquée au Dépôt Légal (DL) et au Département des Archives Professionnelles (DA). Cependant, le DA s'intéressant non pas à la diffusion elle-même du document mais au document lui-même, ces données vont être transformées et complétées de manière à être plus précises et à avoir une meilleure description. Cette description plus fine permet la vente des extraits.

De même que pour le DL, le DA possède plusieurs bases de données partageant les mêmes référentiels :

- la base DAV (Archives Professionnelles de la Télévision Nationale)
- la base DAVREG (Archives Professionnelles de la Télévision Régionale)
- la base DAVRAD (Archives Professionnelles de la Radio)

Ces trois bases de données sont appuyées par plusieurs lexiques et *thesauri*, notamment celui des noms communs⁷ et des personnes physiques et morales.

3.2.3 La base de données juridique (DJ)

La base de données « Adaje » de la DJ comprend l'ensemble des données permettant d'identifier et de rémunérer les ouvrants-droit⁸ et les ayants-droit⁹ des documents et extraits vendus. Cette base juridique contient par conséquent des tables de Personnes, de Contributions, d'Informations personnelles,

7. L'importance — et la complexité — de ce thésaurus au DA nécessite une interface graphique, « Totem », pour le visualiser et cataloguer les documents. Un exemple de visualisation de ce thésaurus est possible en Annexe E : Le thésaurus de noms communs de l'INA.

8. Personnes auxquelles les droits ont été ouverts, le producteur lui-même ou ses ayants-droit.

9. « Un ayant droit est une personne ayant acquis un droit d'une autre personne » in <https://droit-finances.commentcamarche.com/faq/4010-ayant-droit-definition>.

Les bases DL et DA, et celle de la DJ n'ont aucun lien entre elles, mais leur données semblent redondantes notamment pour les personnes physiques et morales. Le projet du *Lac de données*¹⁰ devra permettre l'alignement de ces bases entre elles en évitant les doublons : la base de la DJ enrichira notamment les concepts de personnes physiques et morales déjà créés à partir des données de la DDCOL.

Plusieurs référentiels, parfois similaires, sont présents dans les bases de la DDCOL¹¹ et la DJ présentées ici. Leur structure¹² est différente selon les usages qui ont conduit à leur création, et aux besoins qui en résultent : des notes qualité décrivant la fonction précise des personnes sont présentes dans le lexique des personnes de la DDCOL alors que seul un domaine d'activité général est conservé à la DJ. Les systèmes documentaire et juridique de l'INA ne sont pas interopérables et n'ont pas été conçus pour l'être : d'un côté, soit l'événement de diffusion est prioritaire, soit l'extrait documentaire l'est ; dans l'autre l'information juridique joue ce rôle. Les usages sont tous différents et dirigent le stockage des données dans l'Institut.

3.3 Multiplication des sources de données et des référentiels

De manière à améliorer et enrichir ses données, à faciliter le travail de catalogage, de description et d'indexation, l'INA récupère des métadonnées et des données à l'extérieur auprès de plusieurs fournisseurs. Certains fournisseurs deviennent alors eux-mêmes des référentiels, dont l'identifiant qu'ils fournissent est présent dans les bases de données de l'INA aux côtés des données fournis.

Ainsi, l'INA reçoit des informations concernant les chaînes de provenance, les noms du générique avec les titres, les audiences et le public cible du document¹³, ou encore les grilles de diffusion prévisionnelles et réelles. L'ensemble de ces informations permet d'accompagner la tâche de catalogage en fournissant des champs préremplis. Les fournisseurs de ces données sont multiples¹⁴ et fournissent des données tant sur les programmes que sur les producteurs eux-mêmes :

10. Ce projet est évoqué au Chapitre 8 : Le *Lac de données* de l'INA : le référentiel au centre du modèle

11. Voir Annexe D : Les bases de données de la DDCOL de l'INA (Figure D.1 : Les bases de données de la DDCOL de l'INA).

12. Ces structures sont détaillées dans les chapitres consacrés aux alignements des données de l'INA.

13. Voir Annexe H : Les fournisseurs externes de données de l'INA (Table H.1 : L'apport des données de Médiamétrie dans la description effectuée au DL).

14. Voir Annexe H : Les fournisseurs externes de données de l'INA (Figure H.1 : Les fournisseurs extérieurs de données de l'INA).

- Les données prévisionnelles de diffusion de la télévision sont achetées auprès de la société Plurimédia¹⁵. Les fictions, les documentaires, les dessins animés, les émissions de toutes natures, les magazines, ...sont ainsi décrits au préalable par cette société.
- Les données réelles de la diffusion télévisuelle et radio — date, horaires, parts d'audience, public — sont fournies par Médiamétrie¹⁶, en complément des données —programmation, diffusion, description des contenus — reçues de la part des diffuseurs eux-mêmes.
- Des informations complémentaires sur les programmes sont acquises auprès d'agences de presse comme Kantarmédia¹⁷ ; pour les producteurs, les informations sont obtenues depuis la société Karl More Productions France.

La masse des données conservées à l'INA, ainsi que l'évolution de la législation concernant les documents à conserver, a créé un système documentaire pluriel, aux bases de données éclatées et aux référentiels internes divers. L'interopérabilité n'est pas recherchée et chaque base de données ou référentiel répond à un besoin spécifique — conservation, commercialisation, juridique — d'un département ainsi qu'à un usage particulier — apporter des informations biographiques sur une personne, retrouver cette personne pour la rémunérer, ...

15. Voir <http://www.plurimedia.fr/>.

16. Voir <https://www.mediametrie.fr/>.

17. Voir <https://www.kantarmedia.com/fr>.

Le référentiel, tel que présenté dans les chapitres précédents, n'a pour destination que l'institution qui l'a créé, dans un unique but qui est de répondre à ses propres besoins selon ses activités. Cependant, nous l'avons évoqué, une même institution peut disposer de plusieurs référentiels, parfois similaires mais structurés différemment. Ces référentiels prennent la forme de liste de termes, ou de *thesauri*, qui offrent des clés et des termes normalisés aux documents décrits. L'interopérabilité entre les bases de données n'est pas recherchée, comme celle avec des référentiels externes : le lien n'apparaît pas encore comme essentiel.

Ainsi, le référentiel comme fournisseur de clés et de termes contrôlés n'a qu'un usage interne et spécifique ; il n'est pas utilisable autre part comme le montre la Figure 3.1 : Utilisation principale des référentiels conçus comme fournisseurs de clés.

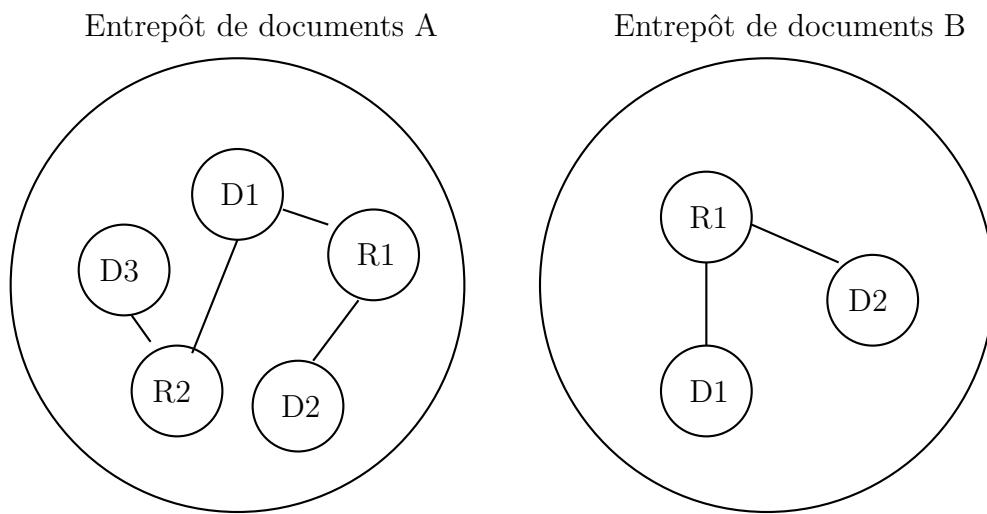


FIGURE 3.1 – Utilisation principale des référentiels conçus comme fournisseurs de clés (R : Référentiel ; D : Document)

Deuxième partie

**RELIER. Vers le partage de
référentiels communs (début des
années 2000 – milieu des années
2010)**

Avec la naissance du Web à la fin du XX^e siècle, le référentiel voit ses usages étendus et multipliés ; sa place en est modifiée. Les différents types de référentiels ont du s'adapter aux nouvelles technologies qui ont alors été offertes. Cependant, ces nouvelles technologies, accompagnées de nouveaux formats, de nouveaux standards et de nouveaux protocoles, ont conduit à la progressive disparition de la notion de référentiel : le référentiel a été divisé en données. Le lien entre ces données devient alors essentiel afin de leur (re)donner du sens entre elles. Mais ces liens peuvent ne pas être seulement présents pour établir une liaison entre deux données : ils peuvent permettre de relier deux jeux de données, et notamment un jeu de données d'une institution avec celui d'une autre. Si le lien n'était présent que dans les *thesauri* pour définir le type de relations, il est désormais l'enjeu principal de l'utilisation d'un autre référentiel ou jeu de données : il permet l'enrichissement de ses propres données avec des données externes.

Le Web de données a permis cet éclatement du référentiel et des jeux de données, et , par sa structure, a nécessité la création de ces liens. Cependant, un type de référentiel, l'ontologie, est toujours indispensable sur le Web car il offre un vocabulaire pour la description du monde réel. L'ontologie permet aussi l'établissement d'un grand nombre de liens entre les données, et permet à une institution de n'utiliser plus un, mais autant de référentiels qu'elle le souhaite.

C'est pourquoi le milieu bibliothéconomique s'est très tôt intéressé au Web de données et participe aux nombreuses réflexions qui l'animent. Les référentiels sont partagés sur ce Web de données et sont repris par d'autres institutions. Si le lien entre les institutions n'est pas l'enjeu principal de la réutilisation de ces référentiels, il est néanmoins créé, et permet la naissance de référentiels de rang supérieur, aggrégateurs de données.

Chapitre 4

Le web de données : une exposition commune des référentiels

« Le web de données, en proposant une forme d’interopérabilité basée sur des standards du Web et sur des liens entre les ressources, semble à même de faciliter l’accès à des données structurées, stockées dans des bases telles que les catalogues de bibliothèques, les inventaires d’archives ou les bases culturelles des musées.¹ »

Le domaine bibliothéconomique, et plus généralement celui culturel, est l’un des premiers à s’être intégré dans le Web de données. Les avantages apportés par le Web, tels que le partage et la mise en commun de référentiels et de données, ont permis une large adoption des standards et formats du Web de données dans les institutions patrimoniales. Cependant, les pratiques individualistes² qui étaient celles des institutions auparavant se retrouvent dans le développement de ce web de données et ont conduit à une efficacité limitée lors de ses débuts.

Malgré ces difficultés des premiers temps, les institutions patrimoniales se sont désormais emparées de ce web de données, devenu un lieu de partage de liens et un fournisseur d’identifiants que les institutions peuvent stocker en vue d’enrichir leurs propres données³.

Plus encore que le partage de liens, le web de données est également un apport

1. S. Dalbin, E. Bermès, A. Isaac, *et al.*, “Approches documentaires...”, p.45.

2. En 1991, Thomas R. GRUBER fait remarquer que ces « blocs monolithiques » propres à chaque institutions sont un frein au développement de liens avec d’autres institutions. « Que peut-il être fait pour permettre l’accumulation, le partage et la réutilisation des bases de connaissances ? ». Voir Thomas R. Gruber, “The Role of Common Ontology in Achieving Sharable, Reusable Knowledge Bases”, dans *Principles of Knowledge Representation and Reasoning : Proceedings of the Second International Conference*, 1991, p. 601-602, URL : <http://citeseer.ist.psu.edu/viewdoc/download?doi=10.1.1.35.1743&rep=rep1&type=pdf> (visité le 09/09/2020), p.1.

3. Un premier exemple a été étudié précédemment avec l’achat de ressources extérieures à l’INA. Voir section 3.3 : Multiplication des sources de données et des référentiels.

considérable dans l'expérience de l'utilisateur qui recherche des données spécifiques sans savoir vers quelle institution se tourner. Il peut, avec les technologies du Web, naviguer de lien en lien, d'institution en institution, rebondir de document en document, sans se rendre compte des frontières techniques ou institutionnelles⁴. Le Web permet un affranchissement des frontières, à la fois pour l'utilisateur final que pour les machines.

Enfin, les technologies du Web, utilisées dans le web de données, proposent de nouveaux formats et de nouvelles modélisations de données, conduisant à la disparition progressive de la notion de référentiel. De plus, la massification des données du web de données posent la problématique de leur accès rapide pour l'utilisateur.

4.1 Le web de données : naissance et principes

La recherche d'un protocole et d'un format d'échange de données entre les institutions est constante. Nous avons évoqué précédemment⁵ les difficultés rencontrées avec les protocoles Z039-50 et OAI-PMH. Ces derniers sont insuffisants pour permettre un partage massif de données et de référentiels, mais ils ont permis l'évolution de la réflexion sur l'interopérabilité. Le grand bouleversement est survenu au milieu des années 2000 avec le Web de données qui ne créé pas de protocole nouveau, mais en utilise un largement répandu et utilisé, HyperText Transfer Protocol (HTTP). Les règles édictées ont permis sa bonne utilisation pour créer le web de données et la seule naissance d'un format d'échange, Resource Description Framework (RDF).

4.1.1 Créer un modèle de données nativement compatible avec le web : le web de données

Naissance du Web de données

Dès 1989, Tim BERNERS-LEE propose un « espace d'information commun »⁶ où les textes seraient liés par des liens⁷. Il propose ainsi un modèle de noeuds et de liens qui permet d'entrer n'importe quel type d'informations, de manière, grâce aux liens, à trouver une ressource sans avoir eu à la chercher. Cependant, plusieurs difficultés s'offrent encore : il est nécessaire, comme dans un arbre, d'avoir des noeuds uniques ; la modélisation du

4. « Sur le Web, un utilisateur a la possibilité de naviguer d'un site à un autre sans avoir connaissance des moyens techniques utilisés pour publier les données, ni même avoir conscience des ruptures ou des frontières entre chacun des sites. » in S. Dalbin, E. Bermès, A. Isaac, *et al.*, “Approches documentaires...”, p.45.

5. Voir section 2.2 : Le *thesaurus*, vocabulaire contrôlé hiérarchique le plus fréquent.

6. « poll of information » in Tim Berners-Lee, *Information Management : A Proposal*, w3, 1989, URL : <https://www.w3.org/History/1989/proposal.html> (visité le 01/08/2020).

7. « a web of notes with links » in *Ibid.*

monde réel est impossible, et le modèle de noeuds et de liens se heurte aux mêmes réflexions que Porphyre et les encyclopédistes quant à la possibilité de représenter le monde en un seul arbre. Pour y faire face, Tim BERNERS-LEE propose le lien hypertexte comme solution.

En 1994, Tim BERNERS-LEE continue la réflexion sur le Web et les liens hypertexte⁸. En 1989, seuls les documents étaient mis sur le Web et Tim BERNERS-LEE décrivait la manière de les relier entre eux. En 1994, il propose d'intégrer au Web des données du monde réel qui seraient reliées par des liens hypertexte, comme les documents, toujours sous la forme de noeuds et de liens. Cette proposition débute la réflexion sur le Web de données. Cependant, la réalité n'est pas compréhensible d'une machine, et les liens ne peuvent se comprendre que par leur contexte : l'ajout de valeurs aux relations permet de donner du sens au Web — c'est le Web sémantique. Tim BERNERS-LEE se fait ainsi le promoteur de la donnée structurée à la fois pour la machine et pour l'humain.

Une feuille de route est par conséquent écrite par Tim BERNERS-LEE en 1998⁹. Il y évoque pour la première fois le terme « web de données »¹⁰, mais la feuille de route n'est pas appliquée et il faut attendre 2006 et la publication *Linked data*¹¹ pour que les recommandations du Web sémantique soient expliquées et adoptées. Fondamentale, cette publication évoque les principes du Web de données actuel en décrivant les bonnes pratiques à adopter. Le Web est ainsi perçu comme une « base de données globale »¹² où les données sont reliées de la même manière que les documents HTML avec des liens hypertexte.

Cette interopérabilité basée sur les liens, théorisée notamment par Tim BERNERS-LEE, est à l'origine du web de données et de l'actuel partage de données et de référentiels entre les institutions patrimoniales. Le lien apparaît comme essentiel et permet de décloisonner chaque institution pour les faire communiquer ensemble de manière à améliorer la recherche de données et de documents par l'utilisateur final¹³.

8. Id., *Plenary talk at WorlWideWeb*, w3, 1994, URL : <https://www.w3.org/Talks/WWW94Tim/> (visité le 01/08/2020).

9. Id., *Semantic Web roadmap*, 14 oct. 1998, URL : <https://www.w3.org/DesignIssues/Semantic.html> (visité le 08/09/2020).

10. « web of data » in *Ibid.*

11. Id., *Linked Data*, 27 juil. 2006, URL : <https://www.w3.org/DesignIssues/LinkedData.html> (visité le 08/09/2020).

12. **bermes_2_2013.**

13. L'utilisateur final n'est pas seulement le grand public, il peut être chercheur, professionnel dans une institution, consommateur commercial, ...

Principes généraux

La publication de 2006 de Tim BERNERS-LEE décrit précisément les principes du web de données qu'il est nécessaire de développer pour comprendre l'évolution des pratiques documentaires des institutions depuis le milieu des années 2000. Ces principes s'appuient sur l'architecture du Web existant et ne visent pas la création d'un Web : l'interopérabilité des données doit passer par une interopérabilité des protocoles et des formats utilisés avec ceux du Web qui connaît une utilisation croissante en 2006.

Le premier principe évoqué est celui de l'utilisation des Uniform Resource Identifier (URI) comme clé unique d'une ressource : en cas de non utilisation du standard des URIs, le Web sémantique n'est plus possible¹⁴. En effet, une URI possède une syntaxe précise qu'il convient de respecter et d'adopter : schème :autorité/chaîne_de_caractères¹⁵.

Le second principe est celui de l'utilisation du protocole du WorldWideWeb, HTTP.

Le troisième principe impose le renvoi d'informations et de données dans des formats standards du Web, en RDF/XML, ou en N3 ou Turtle¹⁶. Tous ces formats admettent le langage de requête SPARQL.

Enfin, le quatrième principe est celui de la création de liens entre les ressources — donc les URIs — sans lesquels les efforts réalisés avec les trois premiers principes sont vains. Ainsi, un web fiable, sans frontières est créé¹⁷ ; l'utilisateur peut y naviguer facilement grâce aux liens hypertextes. Le web de données est par conséquent moins une base de données qu'un lieu où les liens donnent de la valeur aux ressources liées : plus une ressource possède de liens, plus celle-ci a une description précise et fiable, plus elle devient visible à l'utilisateur.

Nous l'aurons remarqué, depuis le début de cette description du Web de données, la notion de référentiel semble d'estomper au profit de ressources et de données liées. En effet, un référentiel n'est qu'une mise en forme spécifique d'un jeu de données selon une structure propre à son producteur. Cette spécificité de chaque jeu de données n'est pas valable dans le Web de données : un retour à la donnée est nécessaire, les liens qui lui seront affectés permettront alors de représenter son ancienne structure dans le référentiel. Tim BERNERS-LEE décrit cette nécessité de se dégager de ses propres formats sur le Web pour évoluer vers des formats compréhensibles par une machine : il crée l'échelle des cinq

14. « If it doesn't use the universal URI set of symbols, we don't call it Semantic Web » in *Ibid*.

15. **bermes_2_2013**.

16. *Ibid*.

17. « serious, unbounded web in which one can find al kinds of things, just as on the hypertext web we have managed to build » in *Ibid*.

étoiles dans ce but, le RDF étant la meilleure des solutions d'exposition des données.

4.1.2 Inventer un format d'échange compatible avec ce modèle de données : RDF

HTTP est le protocole utilisé pour le Web de données ; le format d'échange est RDF. C'est un standard développé pour le Web capable d'assurer l'interopérabilité des données. Seules des URIs peuvent constituer des ressources. Ces ressources sont ensuite reliées grâce à un lien typé par le formalisme offert par RDF. RDF n'offre pas, comme cela est le cas avec les encodages XML archivistiques ou codicologiques, un schéma prédéfini, mais un modèle logique de description des ressources.

RDF ne permet pas de relier directement deux ressources, seulement de typer leur relation afin que la machine puisse interpréter la nature de leur lien, peut importe la localisation des deux ressources. Ainsi, la forme d'un triplet RDF reflète cette distinction : le « sujet » est nécessairement une ressource — par conséquent une URI —, il est suivi d'un « prédicat » qui définit la nature de la relation avec le troisième élément du triplet, l'« objet » qui peut être une ressource ou un littéral. Le triplet est donc une simple phrase sujet-verbe-complément compréhensible par une machine. Une ressource pouvant être à la fois sujet dans un triplet, prédicat dans un autre, ou objet dans d'autres, un graphe se construit alors. L'information est donc totalement déconstruite pour un humain, mais elle devient compréhensible par une machine, qui permet ensuite la reconstruction de l'information par des requêtes efficaces sur ces triplets — cette reconstruction pouvant être personnalisée selon la requête effectuée.

Avec l'apparition du Web de données, un changement d'échelle des référentiels a lieu : ils cessent d'être utilisés par leur seul créateur dès lors qu'ils sont transformés puis envoyés dans le Web de données, ils peuvent désormais être partagés et réutilisés grâce aux URIs. L'utilisation d'un protocole existant, ainsi que d'un nouveau format d'échange, a permis de s'éloigner des modèles d'interopérabilité par conversion et copie, ou par le plus petit dénominateur commun : les référentiels sont des nœuds autour desquels les jeux de données sont rattachés¹⁸.

18. C'est l'interopérabilité de la « roue et de l'essieu », ou « hub and spoke », décrite dans **bermes_2_2013**. Voir Annexe B : Les différents types d'interopérabilité (Figure B.3 : L'interopérabilité de la roue et de l'essieu)

4.2 La mise en commun de référentiels au service des institutions

« Le travail d’alignement, c’est-à-dire de mise en relation, des référentiels entre eux dans l’objectif de créer du lien (et donc de l’interopérabilité) entre les bases et au-delà entre les institutions, initié dans le cadre de la réflexion autour du Web de données, va se poursuivre pour faciliter le maintien du référentiel et son enrichissement.¹⁹ »

La mise en commun de référentiels est essentielle pour créer de l’interopérabilité. Une mise en commun de référentiels au sein même d’une institution est possible de manière à créer du lien entre ses données, mais l’utilisation de plusieurs référentiels externes permet souvent d’exprimer plus de relations et de propriétés que la simple utilisation de référentiels internes.

C'est pourquoi les institutions patrimoniales et culturelles se sont engagées dans le Web de données très tôt et ont permis l'émergence de référentiels internationaux qui font aujourd'hui autorité. Pour accroître encore la puissance de ces référentiels, des passerelles sont créées entre les référentiels, de manière à créer plus de liens.

4.2.1 L'adoption du Web de données en institutions patrimoniales

Tim BERNERS-LEE, dans sa feuille de route pour le Web sémantique en 1998, plaide pour le Web de données. Seulement, sa publication de 2006 rappelle les avantages de ce Web de données ainsi que les pratiques qui y sont liées. En effet, ces technologies étant nouvelles au début des années 2000, elles sont utilisées principalement pour de la recherche : les pratiques du Web sémantique sont par conséquent individuelles et peut conformes aux recommandations de Tim BERNERS-LEE pour le Web sémantique. La finalité n'étant pas la publication sur le Web, ces données sont peu exploitables et parfois non accessibles²⁰.

À la suite des travaux d'un groupe de travail du W3C, le Semantic Web Education and Outreach Interest Group (SWEO)²¹, destiné à promouvoir les technologies du Web

19. Gautier Poupeau, *Réflexions et questions autour du Web sémantique*, Les Petites Cases, 6 oct. 2018, URL : <http://www.lespetitescases.net/reflexions-et-questions-autour-du-web-semantique> (visité le 01/08/2020).

20. Tim BERNERS-LEE dans la publication de 2006 se plaint de cette production conséquente de triplets non accessibles dans le Web sémantique : « Many research and evaluation projects in the few years of the Semantic Web technologies produced ontologies, and significant data stores, but the data, if available at all, is buried in a zip archive somewhere, rather than being accessible on the web as linked data. ». Voir T. Berners-Lee, *Linked Data...*

21. <https://www.w3.org/blog/SWEO/>

sémantique, l'initiative « Linking Open Data »²² naît pour encourager à la publication de données dépourvues de droits. La publication de DBpedia²³ a permis, à partir des pages Wikipédia, de créer des triplets RDF pouvant servir de support aux initiatives des institutions²⁴ et d'améliorer la qualité des triplets existants en créant de nouveaux liens.

Quatre ans après la mise en œuvre du Linked Open Data, le nombre de jeux de données disponibles a été multiplié par vingt²⁵. Pour évaluer l'efficacité du Web de données et les perspectives à venir dans le milieu bibliothéconomique, un nouveau groupe de travail du W3C est lancé en 2010, le *Library Linked Data Incubator Group* (LLD-XG). dans son rapport final²⁶, le groupe préconise d'accentuer encore la coopération entre les institutions, et de faire participer davantage les bibliothèques dans la réflexion du Web de données et l'élaboration de nouveaux standards. Aujourd'hui, l'essor du Web de données en bibliothèque se réalise autour de référentiels faisant autorité.

Le groupe LLD-XG a permis aux institutions patrimoniales de se concentrer davantage sur la publication des données. Ainsi, la Bibliothèque nationale de France (BNF) inaugure la plateforme Data BNF en 2011 pour ouvrir son catalogue ainsi que les données d'autorité au format RDF.

4.2.2 Utiliser des vocabulaires de valeurs

Au-delà de l'interopérabilité souhaitée entre les institutions patrimoniales, les référentiels publiés dans le Web sémantique permettent une utilisation dans des domaines différents : l'utilisateur peut ainsi utiliser plusieurs référentiels du Web de données pour décrire ses données. Ces référentiels deviennent des référentiels de valeurs, compris comme un « ensemble de termes organisés en système de connaissance »²⁷. Ils sont l'essieu de cette interopérabilité de la roue et de l'essieu. Les autorités LCSH de la Library of Congress

22. Cette initiative est aujourd'hui omniprésente dans le Web de données : en juillet 2020, 1260 jeux de données sont présents dans le *Linked Open Data Cloud*. Leur représentation graphique, guidée par les liens entre ces jeux de données, devient au fil des années un exercice de plus en plus complexe face à l'augmentation constante des jeux de données présents. Voir <https://www.lod-cloud.net/>. Voir Annexe I : La constellation du Linked Open Data

23. fr.dbpedia.org

24. E. Bermès, A. Isaac et G. Poupeau, “Convergence et interopérabilité : vers le Web de données”, dans *Le Web sémantique en bibliothèque*, ISSN : 0184-0886, 2013, p. 29-46, URL : <https://www-cairn-info.proxy.chartes.psl.eu/le-web-semantique-en-bibliotheque--9782765414179-page-29.htm> (visité le 01/08/2020).

25. 12 en 2007, 203 en 2010 : voir <https://www.lod-cloud.net/>.

26. Thomas Baker, E. Bermès, Karen Coyle, Gordon Dunsire, A. Isaac, Peter Murray, Michael Panzer, Jodi Schneider, Ross Singer, Ed Summers, *et al.*, *Rapport final du groupe d'incubation "Bibliothèques et web de données" (LLD-XG)*, W3C, 2012, URL : http://mediatheque.cite-musique.fr/MediaComposite/ARTICLES/W3C/XGR-lld-fr.html#xd_co_f=ZWZhZWJmOGEtZjhkMS000DI1LWI4M2YtNTQyZmYxMzg3MzZ1~ (visité le 08/09/2020).

27. S. Dalbin, E. Bermès, A. Isaac, *et al.*, “Approches documentaires...”, p.47.

sont un de ces référentiels de valeurs²⁸. RAMEAU, les autorités de la BNF, sont créées à partir de LCSH et des données de la BNF. Elles permettent d'être utilisées dans plusieurs catalogues, celui de la BNF, mais également celui du Système universitaire de documentation (SUDOC).

À partir d'un seul référentiel commun et partagé, que chaque utilisateur — institution — peut mettre à jour, plusieurs jeux de données peuvent être décrits et indexés. Cependant, la publication de référentiels dans le Web de données n'est pas la priorité des institutions ni leur objectif initial ; cette publication n'intervient qu'après l'opération de catalogage qui aura nécessité la création de nouvelles vedettes.

4.2.3 Créer des passerelles entre les référentiels

Le parcours de liens de ressources en ressources, ainsi que l'alignement des référentiels entre eux, permet la création de passerelles et un enrichissement infini de chaque référentiel. Le lien, une nouvelle fois, est essentiel.

D'abord, le parcours de liens permet des rebonds entre les référentiels. Cette interopérabilité par parcours de liens²⁹ conduit à la découverte de nouvelles ressources que l'utilisateur n'aurait pas trouvées de lui-même. Ainsi, les vedettes LCSH renvoient vers les vedettes identiques ou similaires d'autres référentiels, tels que RAMEAU ou OCLC. De même que pour les liens entre les vedettes de LCSH selon le type de relations, ces liens externes sont également séparés selon la relation de la vedette avec les vedettes visées par les liens³⁰.

Ensuite, des fichiers d'autorité sont nés d'alignements avec d'autres fichiers d'autorités. En effet, la redondance de certaines autorités dans plusieurs référentiels n'est pas opportune dans le Web de données : cela crée de la dissonance et empêche la naissance d'une autorité globale regroupant l'ensemble des informations et des données des autorités existantes. Dans ce but, plusieurs fichiers d'autorité comme le Virtual International Authority File (VIAF)³¹ fusionnent les fichiers d'autorités de bibliothèques nationales du monde entier. Ce projet a été initié dès 2003 par la Library of Congress, la Deutsche Nationalbibliothek, la Bibliothèque nationale de France et OCLC Research³², et compte

28. Voir section 1.1 : Du langage libre au langage contrôlé : vers l'indexation.

29. « follow your nose » in E. Bermès, A. Isaac et G. Poupeau, “Convergence et interopérabilité : vers le Web de données”... Voir Annexe B : Les différents types d'interopérabilité (Figure B.4 : L'interopérabilité par parcours de liens).

30. Voir Figure 4.1 : Modélisation des liens vers des référentiels externes présents dans la vedette « television » de LCSH.

USDA : the National Agricultural Library's Agricultural Thesaurus. Voir <https://agclass.nal.usda.gov/>.

YSO : Yleinen suomalainen ontologia. Voir <https://finto.fi/ys0/fi/>

31. <http://viaf.org/>

32. Id., “Les promesses du Web de données en bibliothèque”, dans *Le Web sémantique en bibliothèque*,

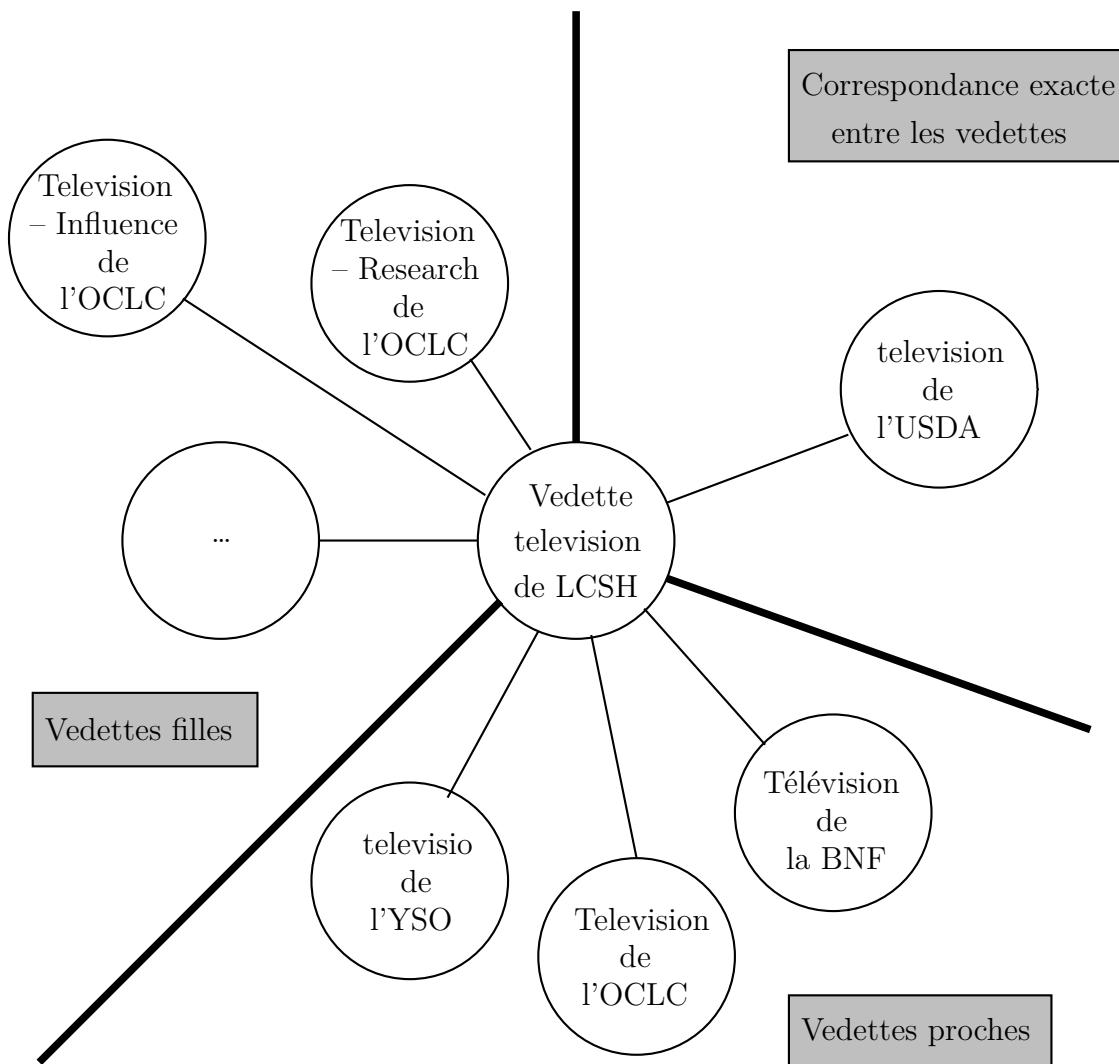


FIGURE 4.1 – Modélisation des liens vers des référentiels externes présents dans la vedette « television » de LCSH

aujourd’hui plusieurs dizaines de bibliothèques partenaires.

L’agrégation de multiples vedettes d’autorités identiques provenant de diverses institutions permet la création d’une *super* fiche d’autorité dans VIAF, créée à partir de liens et n’affichant que des liens³³. En raison de l’origine bibliothéconomique de VIAF, les formes retenues sont exprimées avec leur code MARC 100 ou 200 et celles similaires avec le code 400, de même que les sujets qui y sont liés avec les codes 5XX.

ISSN : 0184-0886, 2013, p. 47-65, URL : <https://www-cairn-info.proxy.chartes.psl.eu/le-web-semantique-en-bibliotheque--9782765414179-page-47.htm> (visité le 01/08/2020).

33. L’autorité personne de Jean-Luc GODARD montre bien cette structuration des notices de VIAF : un graphe permet la modélisation des multiples institutions de récupération des notices d’autorité, et les thèmes et sujets liés à Jean-Luc GODARD sont listés dans la suite de la page de cette vedette.

Les référentiels propres à chaque institution ont été transformés de manière à pouvoir être intégrés au Web sémantique. Avec cette augmentation de ressources, identifiées par des URIs et échangées par RDF, de nombreux liens ont pu être créés entre les référentiels. Les données et les autorités ont ainsi été partagé et quelques référentiels jouent désormais un rôle central dans le Web de données grâce à leur taille et aux nombre de liens qui y renvoient, ou qu'ils renvoient.

4.3 Vers la fin de la notion de référentiels ?

L'éclatement du document en données sur le Web a permis de grandes avancées pour les institutions patrimoniales qui partagent non plus des notices bibliographiques ou d'autorités, mais des données liées au travers d'URIs. Elles ont trouvé avec le Web de données un protocole ainsi qu'un format d'échange standardisés et utilisés par tous les utilisateurs. Cependant, ce règne de la donnée sur le Web conduit à de nouvelles réflexions quant à la définition des référentiels : personnes, lieux et sujets sont considérés comme des données de référence ; pourquoi alors ne pas considérer une œuvre comme une donnée de référence elle aussi ? Cette conceptualisation de la réalité conduit ainsi à repenser les modèles de données dans les institutions ainsi que les formats de description des documents afin de partager des formats et des standards sur le Web pour profiter au plus grand nombre.

4.3.1 Quand tout devient un potentiel référentiel

D'un référentiel qui était une liste de mots contrôlés et hiérarchisés avec les *thesauri*, le Web de données le fait passer vers une nouvelle définition et de nouveaux usages : un référentiel est désormais un ensemble d'informations qui sont susceptibles d'être partagées puis réutilisées dans divers systèmes documentaires afin de créer du lien³⁴. Ces informations ne sont plus spécifiquement des termes choisis et contrôlés par des documentalistes : tout peut devenir information et par conséquent référentiel s'il y a des relations établies.

Dès la fin du XX^{ème} siècle, deux modèles voient le jour pour repenser la structure de la donnée et la place des référentiels. En 1996, la réflexion autour de la description des collections muséales aboutit à la création du modèle du Comité International pour la Documentation - Conceptual Reference Model (CIDOC-CRM)³⁵ qui est un modèle orienté document – objet — permettant de pouvoir décrire les interactions de chaque objet avec d'autres entités.

34. *Ibid.*, §49.

35. CIDOC-CRM, URL : <http://www.cidoc-crm.org/> (visité le 09/09/2020).

En parallèle de ce modèle destiné aux descriptions de collections de musées, les bibliothèques mènent également une réflexion similaire entre 1992 et 1997, ce qui conduit à l'élaboration des modèles Functional Requirements for Bibliographic Records (FRBR)³⁶. Il est nécessaire de s'attarder sur ces FRBR afin de mieux comprendre la structure du *Lac de données* de l'INA expliquée par la suite³⁷, bien que celle-ci ne soit pas exactement similaire aux FRBR. Trois groupes sont distingués dans les FRBR³⁸ : l'un correspond à la notice bibliographique elle-même, les deux autres aux points d'accès.

La notice bibliographique est, avec les FRBR, divisé en quatre sections, partant des caractéristiques propres de l'exemplaire décrit — l'item —, suivies par les caractéristiques de la publication auquel le document appartient — la manifestation — et par celles de son contenu — l'expression —, pour terminer avec celles de la création abstraite auquel le document appartient — l'œuvre.

Le premier point d'accès est le groupe des personnes et des collectivités qui permettent de décrire les responsabilités de chacun — auteur, producteur, ayant-droit, ... —, de l'œuvre à l'item. Le second point d'accès permet la description du contenu, le sujet de l'activité intellectuelle ou artistique à travers des concepts, des objets, des événements ou des lieux.

La création de liens entre les différents groupes et les différentes entités permet une description fine des contenus, des responsabilités et des œuvres ; la liaison entre des entités conformément au Web de données ; la création infinie de référentiels avec une œuvre pouvant être sujet d'une autre œuvre par exemple³⁹.

Avec ces nouveaux modèles — CIDOC-CRM et FRBR —, toutes les notions peuvent devenir des référentiels : les œuvres, les expressions, les sujets, les familles, Leur avantage est leur partage désormais possible avec d'autres métiers, et plus largement sur le Web de données puisqu'une exposition en RDF est possible et réalisée. Plusieurs institutions peuvent alors utiliser les données modélisées selon les FRBR afin de créer plus

36. Plusieurs modèles FRBR spécifiques ont vu le jour : les Functional Requirements for Authority Data (FRAD) (pour les données d'autorité ; voir Fédération internationale des associations de bibliothécaires et de bibliothèques, *Fonctionnalités requises des données d'autorité – Rapport final*, 2010, URL : https://multimedia-ext.bnf.fr/pdf/frad_rapport_final.pdf) et les Functional Requirements for Subject Authority Data (FRSAD) (pour les sujets ; voir Id., *Fonctionnalités requises des données d'autorité matières – Rapport final*, 2010, URL : https://multimedia-ext.bnf.fr/pdf/frsad_rapport_final.pdf), destinés à modéliser les relations et les entités des points d'accès.

37. Voir Chapitre 8 : Le *Lac de données* de l'INA : le référentiel au centre du modèle.

38. Le rapport détaillé des FRBR indique l'ensemble des groupes et des relations possibles, ce que nous ne développerons pas ici. Voir Id., *Fonctionnalités requises des notices bibliographiques – Rapport final*, 2012, URL : https://multimedia-ext.bnf.fr/pdf/frbr_rapport_final.pdf

39. Voir Annexe J : Repenser la place du référentiel (Figure J.1 : Le modèle FRBR).

de liens que si elles ne s'étaient appuyées uniquement sur les référentiels communs, tels que LCSH ou RAMEAU, pour établir la description de leur objet.

4.3.2 Vers une uniformisation internationale de la donnée sur le Web et l'adoption de RDF comme format de production

L'apparition des nouveaux modèles centrés sur les entités, dans lesquels toutes les entités sont potentiellement partageables et utilisables par d'autres pour servir de référentiel, favorise de nouvelles réflexions sur le catalogage des documents : la finalité devenant de plus en plus souvent la publication des données sur le Web sémantique, est-il toujours nécessaire de cataloguer dans un format pour ensuite convertir les données en RDF ?

La nécessité d'améliorer les règles de catalogage dans les pays anglo-saxons dans les années 2000 a conduit à l'utilisation des nouveaux modèles FRBR dans le code Resource Description and Access (RDA) publié en 2010. Ce changement majeur est testé à partir de 2011 à la Library of Congress, puis adopté en 2013 dans les pays anglo-saxons ; la France et ses agences bibliographies ABES et BNF tente d'intégrer ces nouvelles règles.

En effet, RDA est nativement pensé autour du Web de données et de l'utilisation en ligne des données de manière à pouvoir ensuite exprimer les entités et leurs relations sous la forme de triplets RDF grâce à l'attribution d'un identifiant à chaque élément ou valeur⁴⁰. Ce code de catalogage a vocation à être internationalement utilisé, de manière à uniformiser les données produites et favoriser ainsi les échanges.

En adoptant RDA, le format MARC est délaissé et semble ne plus pouvoir répondre aux changements impliqués par le Web. Cependant, RDA, publié et utilisé, subit déjà des évolutions et un nouveau modèle de données, nativement centré sur RDF, est créé en 2012 : Bibframe⁴¹. Le format MARC est abandonné au profit d'un modèle pensé autour de RDF, et par conséquent des usages numériques des utilisateurs. Le modèle de données de Bibframe est semblable à celui des FRBR avec les *work*, les *instance* et les *item*. Seulement, RDF n'apparaît plus comme un format de sortie des données après leur catalogage ; il est désormais le format natif de catalogage. Ainsi, chaque donnée, chaque entité, devient un référentiel en ce qu'elle est nativement liée à d'autres ressources.

L'impact du Web de données sur la notion de référentiel s'étend au-delà de la réflexion sur la définition et la structure d'un référentiel : toute la modélisation des données des institutions est remise en question. Ces dernières doivent s'adapter et tenter de trouver

40. Les vocabulaires RDA ont fait l'objet d'un groupe de travail après la création du code de catalogage RDA.

41. Library of Congress, *Overview of the BIBFRAME 2.0 Model*, 2016, URL : <https://www.loc.gov/bibframe/docs/bibframe2-model.html> (visité le 09/09/2020).

de nouveaux modèles de données qui puissent répondre à la fois à leurs besoins internes de description et de signalement des collections, et à la fois aux besoins croissants des utilisateurs sur le Web de pouvoir trouver des ressources en pouvant s'affranchir des barrières technologiques et institutionnelles. Ainsi, il n'y a plus de référentiels dans lesquels l'utilisateur peut aller, les catalogues des institutions ouverts en RDF sont eux-mêmes ces référentiels.

L'apparition du Web et son adoption par le public l'a rendu indispensable pour les institutions et l'obtention d'informations. Les frontières auparavant visibles avec les portails ont été effacées avec le Web de données qui permet la liaison des jeux de données entre eux, peut importe leur provenance et leur lieu de stockage. La réflexion entamée dès le début des années 2000 a permis la réalisation des premiers jeux de données du Web sémantique au milieu de cette décennie. Les possibilités offertes par le Web de données ont conduit l'ensemble des institutions à adopter cette exposition des données, de manière à enrichir leurs propres données avec l'ajout de liens multiples.

L'effacement de la distinction entre document et référentiel conduit à la disparition progressive de la notion de référentiel au profit de nouveaux modèles de données reliant chaque donnée à d'autres à travers un modèle de données partagé, RDF. Cependant, la création de triplets RDF nécessite toujours un certain type de référentiel, l'ontologie. Cette dernière permet d'établir les liens entre les ressources et de les typer.

Chapitre 5

Partager des structurations similaires de jeux de données par les classes et les propriétés : les ontologies, grammaires communes mais spécifiques

L'éclatement des référentiels dans le Web de données conduit à la création, ou au renforcement, de liens entre eux, ainsi qu'entre leur données. Ces liens doivent porter une valeur sémantique de manière à ce que l'éclatement ait lieu sans perdre d'informations. De nouveau, il faut alors que le sens des relations soit contrôlé et partagé par le plus grand nombre : les ontologies permettent cela. Ainsi, si le référentiel présenté sous la forme de liste ou de thésaurus disparaît au profit du Web de données, de nouveaux référentiels, adaptés au Web sémantique, prennent le relais.

Souvent confondues avec les systèmes organisés de connaissances¹, les ontologies diffèrent par leur formalisme. La Partie I : CONTRÔLER. A la recherche de clés (années 1960 – fin des années 1990) a montré qu'une interopérabilité par les référentiels est possible et que les arbres de classification portent difficilement du sens, l'arborescence permettant essentiellement d'effectuer une classification ; avec les ontologies, une interopérabilité sémantique peut avoir lieu.

De même qu'avec le Web de données, le milieu bibliothéconomique a été l'un des premiers à adopter massivement les ontologies afin de pouvoir typer les relations entre les entités : l'ontologie est essentielle au Web sémantique.

1. Ils sont fréquemment nommés KOS pour *Knowledge Organization Systems*.

5.1 L'ontologie, un vocabulaire structurant

Dans le chapitre précédent², nous avons évoqué un premier type de référentiel — les vocabulaires de valeurs — présent dans le Web de données. De manière à pouvoir décrire ces ressources³, d'autres référentiels sont nécessaires, les ontologies, en fournissant les classes et les propriétés utiles aux descriptions. Au-delà de l'apport de ces éléments, l'ontologie permet également une description formelle par des axiomes et des règles de raisonnement, visibles dans le Web de données avec RDF Schema (RDFS) et Web Ontology Language (OWL).

L'ontologie informatique est un concept récent, né à la fin du XX^{ème} siècle comme le Web. Plusieurs types d'ontologies existent, reflétant leur caractère universel ou non, leur domaine de description ; leur structuration et leur formation doivent cependant répondre à des critères précis de manière à structurer le plus efficacement possible de référentiel.

5.1.1 Origines de l'ontologie informatique

« [Les ontologies sont] des vocabulaires de termes — classes, relations, fonctions, constantes d'objet — avec des définitions communes, sous la forme d'un texte compréhensible par les humains et applicable à la machine, de contraintes déclaratives dans leur forme la mieux formée.⁴ »

L'ontologie est d'abord une science philosophique, née avec les *Catégories* d'Aristote, étudiant la réalité des entités, les relations qu'elles entretiennent — hiérarchie, similarité — pour trouver les similarités et les différences présentes dans le monde. Au XIX^{ème} siècle, le siècle de la taxonomie, cette science philosophique devient l'étude de l'ensemble des connaissances existantes dans le monde⁵.

Ce n'est qu'en 1991⁶ puis 1993⁷ que Thomas R. GRUBER, souhaitant améliorer

2. Voir Chapitre 4 : Le web de données : une exposition commune des référentiels.

3. « L'une des fonctions des ontologies est de permettre de définir la nature des ressources » in E. Bermès, A. Isaac et G. Poupeau, “Convergence et interopérabilité : vers le Web de données”..., §49.

4. « vocabularies of representational terms — classes, relations, functions, object constants — with agreed-upon definitions, in the form of human-readable text and machine-enforceable, declarative constraints on their well formed use » in T. R. Gruber, “The Role of Common Ontology in Achieving Sharable, Reusable Knowledge Bases”..., p.2

5. C. Welty et N. Guarino, “Supporting ontological analysis of taxonomic relationships”, *Data & knowledge engineering*—39 (2011), Place : Amsterdam Publisher : Elsevier OCLC : 67323599, p. 51-74.

6. T. R. Gruber, “The Role of Common Ontology in Achieving Sharable, Reusable Knowledge Bases”...

7. Id., *Toward principles for the design of ontologies used for knowledge sharing*. OCLC : 123335010, Stanford, 1993, URL : <https://tomgruber.org/writing/onto-design.pdf> (visité le 09/09/2020).

l'intelligence artificielle et l'indexation structurée, évoque l'ontologie informatique, pensée comme un ensemble d'entités déclaratif destinée au partage des connaissances entre les machines : « Une ontologie est une spécification explicite d'une conceptualisation »⁸. Cette définition donnée très tôt par Thomas R. GRUBER permet d'observer deux principes de l'ontologie : premièrement, elle est une conceptualisation d'un domaine, par conséquent elle est un choix de description sur un domaine précis ; deuxièmement, cette conceptualisation est spécifiée, c'est à dire qu'elle a une description formelle.

Rudi STUDER apporte des précisions en 1998⁹ en proposant une nouvelle définition plus spécifique de l'ontologie : « Une ontologie est une spécification formelle et explicite d'une conceptualisation partagée »¹⁰. Une ontologie est formelle de manière à pouvoir être comprise par une machine ; elle est une spécification explication par la déclarativité de ses concepts, de ses propriétés, ... ; elle est partagée car elle prend l'ensemble des connaissances d'une communauté, d'un domaine ; enfin, la conceptualisation renvoie au domaine décrit par cette ontologie.

L'ontologie est un référentiel de classes et de propriétés, ne s'appliquant qu'à un seul domaine particulier de la connaissance, mais permettant de le structurer. Son fort développement a permis une application dans le Web de données et dans le milieu bibrathéconomique, qui considère les ontologies comme « des éléments de description de métadonnées »¹¹.

5.1.2 Des ontologies différentes

Une grande diversité d'ontologies existe. Certaines sont plus importantes que d'autres du fait du nombre d'utilisations qu'elles entraînent et de leur généralité ; d'autres, plus spécifiques, paraissent moins importantes par le faible nombre de liens qu'elles suscitent. Les ontologies peuvent également être classées selon les usages qui en sont faits, selon si leur finalité est une publication sur le Web sémantique ou simplement une utilisation interne à une institution.

Au plus haut niveau se trouvent des ontologies « noyaux »¹² qui modélisent les connaissances communes, partageables et réutilisables d'un domaine à un autre : le modèle

8. « An ontology is an explicit specification of a conceptualization. » in *Ibid.*, p.1.

9. Rudi Studer, V.Richard Benjamins et Dieter Fensel, “Knowledge engineering : Principles and methods”, *Data & Knowledge Engineering*, 25–1 (1998), OCLC : 4927801743, p. 161-197.

10. « An ontology is a formal, explicit specification of a shared conceptualization » in *Ibid.*

11. T. Baker, E. Bermès, K. Coyle, *et al.*, *Rapport final du groupe d'incubation "Bibliothèques et web de données" (LLD-XG)...*

12. A. Isaac, “Les référentiels : typologie et interopérabilité”, dans, 2012, p. 85-104, URL : <https://hal.inria.fr/hal-00740282> (visité le 01/07/2020), p.4.

CIDOC-CRM des musées propose ainsi une ontologie réutilisée dans RDFS notamment¹³. Son vocabulaire correspond aux événements, aux objets, aux moments, ... ce qui en fait un vocabulaire générique pour d'autres ontologies de plus bas niveau.

Au niveau inférieur se trouvent les ontologies de domaine, propres à un domaine en particulier : elles modélisent les connaissances de ce domaine uniquement ; elles offrent des concepts et des relations permettant de décrire les activités et les vocabulaires du domaine en question. Les concepts de ces ontologies de domaine sont souvent des spécialisations¹⁴ d'ontologies de plus haut niveau. L'ontologie FRBR¹⁵ peut être considérée comme une de ces ontologies de domaine car elle utilise RDFS, les *Dublin Core Terms* (DC Terms), et d'autres ontologies de haut niveau¹⁶.

Plus bas encore, il existe des ontologies utilisées dans l'unique cadre d'utilisations applicatives, par un petit nombre d'utilisateurs. Elles ne modélisent par conséquent que les termes et les relations nécessaires à l'application, en utilisant principalement des ontologies de plus haut niveau, notamment celles de domaine¹⁷.

De multiples ontologies existent et diffèrent de cette hiérarchisation¹⁸ et participent à la diversité des ontologies. Cette diversité est essentielle pour décrire tout type d'objets ou d'événements — que ce soit dans le milieu culturel ou non —, ou pour relier à ces ontologies des vocabulaires propres : elle participe à l'interopérabilité des référentiels entre eux. Cette interopérabilité, grâce aux ontologies, peut être entre un système documentaire interne et le Web de données, ou bien entre deux jeux de données du même Web de données.

5.1.3 Les principes de l'ontologie

Permettre l'interopérabilité entre tout type de données et de jeux de données nécessite une structure et des principes. Dans la publication de 1993¹⁹, Thomas R. GRUBER édicte déjà cinq critères sans lesquels une ontologie ne peut pas être formée correcte-

13. Voir Annexe L : Ontologies de haut et de bas niveaux (Figure L.1 : Une ontologie de haut niveau : CIDOC-CRM).

14. Ainsi que le faisait remarquer Rudi STUDER en 1998 in R. Studer, V. Benjamins et D. Fensel, "Knowledge engineering..." .

15. Décrise dans la section 5.3 : Les ontologies dans le Web sémantique.

16. Voir Annexe L : Ontologies de haut et de bas niveaux (Figure L.2 : Une ontologie de domaine : FRBR)

17. L'ontologie *Citation Counting and Context Characterization Ontology*(C4O) est l'une de ces ontologies applicatives, ayant peu d'ontologies l'utilisant, et utilisant un grand nombre d'ontologies de plus haut niveau. Voir Annexe L : Ontologies de haut et de bas niveaux (Figure L.3 : Une ontologie applicative : C4O).

18. A. Isaac, "Les référentiels...", p.2.

19. T. R. Gruber, *Toward principles for the design of ontologies used for knowledge sharing....*

ment. Ces critères, généraux, contraignent la graphie tout en laissant le champ ouvert aux modifications futures d'une ontologie. Il préconise ainsi :

- la clarté des termes décrits : leur description doit être objective, complète, et réalisée dans le langage naturel ;
- la cohérence des axiomes retenus et l'interdiction de la discordance entre les termes et les axiomes : elle permet la spécialisation de la conceptualisation qui est la définition de l'ontologie ;
- la possibilité d'étendre l'ontologie même après sa création : cela permet d'accepter les changements d'usages ou de besoins liés à cette ontologie, et par conséquent de la faire évoluer facilement ;
- le biais d'encodage doit être minimal de manière à permettre la plus grande interopérabilité ;
- l'engagement dans l'ontologie doit être minimal, les termes utilisés souvent être ceux les plus souvent utilisés : cela permet la réutilisation de l'ontologie

Ces cinq critères ontologiques dirigent la structure et la nature des éléments essentiels aux ontologies et les constituent. Les concepts sont le premier de ces éléments : ils permettent de définir des idées, des objets ou des notions. De même que pour les vocabulaires contrôlés²⁰, plusieurs propriétés peuvent s'appliquer à ces concepts. Nicola GUARINO décrit ces propriétés²¹ de généricté — absence d'extension pour le concept —, d'identité, de rigidité — si une instance du concept reste en permanence son instance —, d'anti-rigidité — si une instance est principalement définie par son appartenance à un autre concept — et d'unité des concepts. Ainsi, deux concepts peuvent être disjoints, équivalents ou dépendants.

Les relations sont une autre partie essentielle des ontologies, sans lesquelles les concepts n'ont pas de sens entre eux et ne peuvent pas être formalisés. Elles peuvent être inclusives — hiérarchiques —, ou ensemblistes avec des unions, des intersections ou des exclusions. À ces relations peuvent d'ajouter des axiomes, des règles, qui viennent régir les contraintes, les relations ou les concepts eux-mêmes.

Les différents principes des ontologies les rendent strictes sur leur formation et leur structure, de nombreuses propriétés s'imposant. Complexes, ils permettent néanmoins la création d'un vocabulaire à la fois contrôlé, hiérarchique et utilisable par tous.

L'ontologie peut alors trouver son intérêt sur le Web avec une indexation réalisée

20. Voir section 1.2 : Une clé entre les données : les vocabulaires contrôlés.

21. Nicola Guarino et FOIS International Conference on Formal Ontology in Information Systems, *Formal ontology in information systems : proceedings of the first international conference (FOIS'98)*, June 6-8, Trento, Italy, Amsterdam ; Tokyo, 1998.

par les moteurs de recherche ; sur le Web et en institution en permettant la description de jeux de données par des ontologies publiques ; en structurant la connaissance du monde. L'une des applications principale est bibliothéconomique avec la possibilité de valoriser et de publier les collections sous forme de métadonnées. Enfin, l'ontologie permet de relier deux jeux de données, deux référentiels, pourtant éloignés, selon un vocabulaire commun partagé publiquement.

5.2 Des *Knowledge Organization Systems* (KOS) à Simple Knowledge Organization System (SKOS) : vers l'interopérabilité syntaxique

Les ontologies ressemblent fortement aux *thesauri* et autres vocabulaires contrôlés — les Knowledge Organization Systems (KOS) — à cause du contrôle de la graphie, des termes retenus ou rejetés, et de l'établissement de relations entre les termes. Cependant, une ontologie n'est pas nécessairement un thésaurus, alors qu'un thésaurus est une ontologie. Bien que la distinction entre les deux soit mince, elle est essentielle en raison du formalisme qui compose les ontologies.

L'interopérabilité sémantique permise par les ontologies bâtit de l'interconnexion entre les jeux de données, rend possible l'échange et la publication de données nativement différentes sur le Web de données. La conversion par les institutions patrimoniales d'une partie ou de l'ensemble de leurs vocabulaires en ontologies a été permise par l'ontologie SKOS.

5.2.1 Distinguer les systèmes d'organisation de la connaissance des ontologies

Les KOS sont des référentiels contrôlés de vedettes et de termes qui ne sont valables que pour un domaine d'activité et de la connaissance. Ils sont souvent organisés par des relations terminologiques et sémantiques qui les font se confondre avec les ontologies en raison de leur formalisation²². Plus encore, sur le Web de données, les KOS ne jouent pas de rôle, ils n'apportent pas de structure, mais complètent des référentiels ou des jeux de données déjà existants. Les KOS n'offrent qu'une succession de termes destinés à remplir des champs descriptifs, alors que les ontologies dirigent directement les données et leur structure.

22. S. Dalbin, E. Bermès, A. Isaac, *et al.*, “Approches documentaires...”, p.48.

Les ontologies permettent de dépasser certaines limites des KOS relevées dans la Partie I : CONTRÔLER. A la recherche de clés (années 1960 – fin des années 1990) ²³ :

- les *thesauri* et autres vocabulaires sont destinés d'abord à un utilisateur humain qui peut facilement comprendre la structure du vocabulaire ²⁴; les ontologies visent quant à elles à d'abord être comprises par une machine et le Web sémantique;
- les KOS induisent souvent des relations diverses et variables, notamment dans le cas de relations génériques-spécifiques

Les limites identifiées sont dues principalement aux relations des termes des KOS : ces vocabulaires ne sont pas sémantiques, seulement hiérarchiques dans le but de classer et d'aider l'humain dans la compréhension du contexte du terme. L'apport des ontologies est l'ajout de sens aux termes grâce aux relations qu'ils entretiennent entre eux.

5.2.2 SKOS : exposer les systèmes d'organisation de la connaissance sur le Web de données

Dans le but de permettre aux institutions d'établir des liens entre leurs référentiels et leurs données, SKOS a été créé. SKOS est « une ontologie qui se veut simple et compatible avec une majorité d'approches d'organisation des connaissances existantes (thèsaurus, classifications...) » ²⁵. Elle permet de représenter presque tous les types de vocabulaires en concepts liés. SKOS n'est pas un référentiel, seulement un moyen de faciliter l'interconnexion entre les données, leur échange, et de créer de nouveaux usages jusqu'alors impossibles.

Pour cela, SKOS, vocabulaire RDF, reprend les propriétés des vocabulaires contrôlés, notamment du thésaurus. En effet, il est possible d'indiquer des termes préférentiels, alternatifs, traduits, variants, ... ; de plus, des notes peuvent être introduites ; enfin, des liens — génériques et hiérarchiques, associatifs, ... — sont créés entre les différents concepts ²⁶. SKOS a le rôle de l'essieu dans le modèle de l'interopérabilité de la roue et de l'essieu. Elle offre un petit nombre de concepts servant à la description et à la transcription d'un thésaurus dans un langage compréhensible par une machine. Ainsi, SKOS est une ontologie de domaine : elle hérite d'autres ontologies comme DC Terms ou RDFS, et sert d'ontologie de haut niveau pour des ontologies de plus bas niveau ²⁷.

23. A. Isaac, “Les référentiels...”.

24. C'est le cas de la visualisation graphique du thésaurus des noms communs de l'INA. Voir Annexe E : Le thésaurus de noms communs de l'INA (Figure E.1 : Extrait du thésaurus de noms communs de l'INA).

25. *Ibid.*, p.8.

26. Voir Figure 5.1 : Modélisation simplifiée de SKOS.

27. La consultation de <https://lov.linkeddata.es/dataset/lov/vocabs/frbr> permet de constater ce rôle pivot de SKOS pour les autres ontologies, ainsi que sa dépendance à d'autres ontologies.

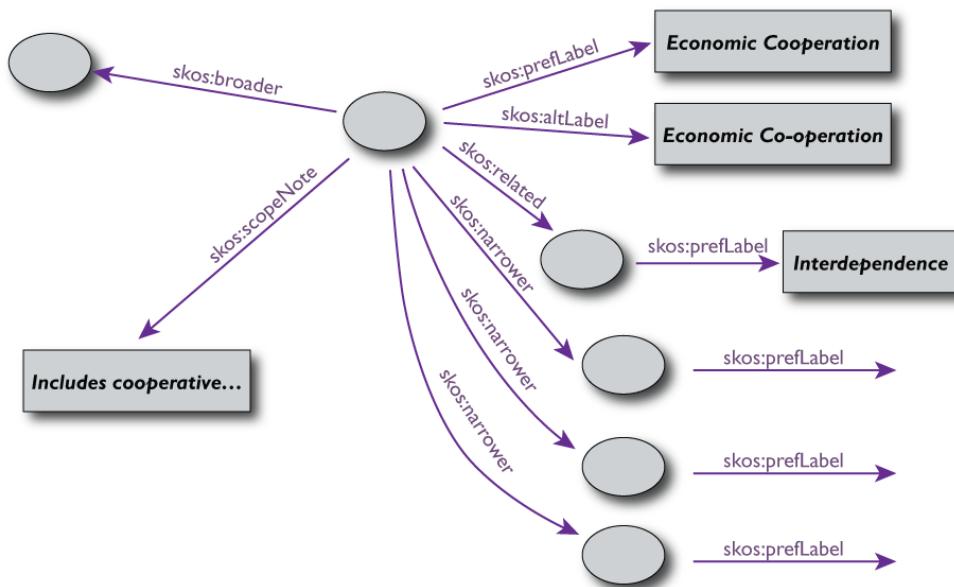


FIGURE 5.1 – Modélisation simplifiée de SKOS [Source : www.w3.org]

Enfin, des propriétés de SKOS permettent de créer du lien entre des concepts provenant de différentes sources : ce sont les propriétés d'équivalence ou de similitude <skos :broader>, <skos :exactMatch> ou <skos :closeMatch>. Ces propriétés permettent le rapprochement de plusieurs concepts : le LCSH²⁸ peut ainsi utiliser SKOS pour créer des fiches de liens. De même, la Bibliothèque nationale de France (BnF) contient des références à l'ontologie Dewey qui lui permet ainsi d'obtenir des relations d'équivalence avec ses données. Alors, l'emploi de SKOS nécessite des URIs de manière à créer des triplets RDF.

Avec SKOS, l'interopérabilité sémantique est désormais possible entre les institutions sur le Web de données. Cette ontologie RDF permet de rapprocher des jeux de données et des référentiels jusqu'alors séparés et pourtant similaires. Le vocabulaire offert pour décrire les *thesauri* et ensuite les partager dans le Web sémantique permet de nombreuses applications en institutions, et facilite ainsi les opérations d'indexation ou de recherche.

Cependant, les vocabulaires des institutions n'étant pas nativement liés entre eux, il reste difficile de les aligner, et par conséquent de passer d'un KOS à une modélisation SKOS, notamment pour les relations partie-tout. Sylvie DALBIN prend en exemple²⁹ ce type de relation qui peut être exprimé dans le Web sémantique par trois types de relations : hiérarchique, instance-classe, ou bien sous-classe-classe. Ces incertitudes rendent

28. Voir Figure 4.1 : Modélisation des liens vers des référentiels externes présents dans la vedette « television » de LCSH.

29. S. Dalbin, E. Bermès, A. Isaac, *et al.*, “Approches documentaires...”.

le processus d'ontologisation complexe, qui l'est d'autant plus quand les termes sont dans des langues différentes.

Les ontologies apparaissent comme un référentiel essentiel dans le Web de données, plus encore que les autorités qui sont devenues des données ; elles ont permis l'apparition d'un Web sémantique. Seulement, la problématique de l'alignement de référentiels entre eux sur le Web de données est toujours présente et ne sera jamais totalement résolue.

5.3 Les ontologies dans le Web sémantique

La finalité principale des ontologies est leur exposition sur le Web de données. L'utilisation qui s'ensuit permet la création d'un Web sémantique, structuré et aux données partagées. Le référentiel compris dans le sens de KOS n'intervient plus autant dans ce Web sémantique ; le modèle de description de ces référentiels s'impose sous la forme des ontologies et améliore dans le même temps la description des concepts qu'il contient par des relations typées.

Cette avancée permet une description plus fine et partagée des documents des institutions patrimoniales : le référentiel est à la fois leur propres données et celles du Web de données, liées par les ontologies publiques.

5.3.1 Décrire des ontologies en RDF : RDFS et OWL

De même que SKOS est une ontologie permettant la description de KOS, RDFS et OWL sont les représentations des ontologies RDF. Les documents décrits par les institutions le sont par des formats et des logiques différentes. Utiliser une seule ontologie peut s'avérer difficile ; en effet, elle peut être trop large ou trop spécifique pour le domaine décrit. C'est pourquoi l'utilisation des URIs, des liens hypertexte du Web, permet d'utiliser autant d'ontologies que nécessaire, et de créer une interopérabilité par parcours de liens, par rebonds sur les URIs.

La constitution de ce réseau de liens utilisé pour la description de documents n'est possible qu'avec l'utilisation du Web sémantique et de RDF. C'est pourquoi il a été nécessaire de construire des modèles de représentation des ontologies en RDF.

RDFS est un langage de description simple, destiné à apporter les bases d'une description en RDF avec la déclaration de classes — et de sous-classes — et de propriétés. Les classes sont des concepts, des types de ressources, identifiés par des URIs. Les propriétés sont les relations qui existent entre les classes. Ainsi, chaque ressource est instanciée à

une classe par la propriété — le prédicat RDF — <<http://www.w3.org/1999/02/22-rdf-syntax-ns#type>> (`rdf:type`). La présence de sous-classes permet de créer des groupes au sein de classes : en déclarant une instance d'une sous-classe, un second triplet est alors implicitement créé entre la sous-classe et la classe avec le prédicat `rdfs:subClassOf`. Tout ce qui s'applique à une classe est appliqué à une sous-classe.

RDFS admet aussi la déclaration de domaines et de codomaines pour les propriétés. Ces propriétés peuvent être de type ressource quand elles relient deux ressources désignées par des URIs, ou bien de type donnée quand l'objet est un littéral. Ce comportement des propriétés est déclaré avec le domaine et le codomaine : le domaine de la propriété définit la type de la classe sujet, tandis que le codomaine définit la classe de la ressource objet — ou du type de donnée si c'est un littéral.

Malgré les classes et sous-classes, et les domaines et codomaines, RDFS reste simple et ne permet pas de description complexe de relations. C'est pourquoi OWL est une extension de RDFS. Des contraintes sur les relations comme la symétrie, l'équivalence, la différence ou la contradiction peuvent être exprimées ; de même, la déclaration d'une liste d'instances contrôlées peut être faite. Comme avec SKOS, il est possible, et souvent indispensable, de déclarer des relations d'équivalences entre les classes, les propriétés ou les instances. Ainsi, une instance peut hériter des propriétés de la classe à laquelle sa classe est équivalente. Pour cela, trois propriétés OWL existent : <<http://www.w3.org/2002/07/owl#equivalentClass>> (`owl:equivalentClass`), <<http://www.w3.org/2002/07/owl#equivalentProperty>> (`owl:equivalentProperty`) et <<http://www.w3.org/2002/07/owl#sameAs>> (`owl:sameAs`).

5.3.2 Utilisation des ontologies en institutions

Chaque ontologie traitant d'un domaine particulier de la connaissance ou du monde, elles sont très nombreuses. Dans sa constante réflexion sur la description, l'indexation et le partage de ses données, le milieu bibliothéconomique s'est emparé du Web de données pour faciliter ses missions et la recherche. Ainsi, plusieurs ontologies sont essentielles dans ce milieu. La première, DC Terms³⁰, et la plus ancienne car créée en 1995, permet la description bibliographique d'un document sur le Web avec quinze propriétés, accompagnées de propriétés affinées — *abstract* l'est de *description*. L'ontologie Dublin Core est constamment utilisée en bibliothèques³¹, et plus généralement dans le Web de données, car elle offre les outils de base servant à la description d'un document.

Pour la description des autorités, l'ontologie FOAF³² est disponible et est également fortement utilisée. Crée au milieu des années 2000, elle permet la description des agents,

30. *Dublin Core Metadata Initiative...*

31. La Figure 5.2 : L'ontologie DC Terms montre la quantité d'ontologies l'utilisant.

32. *FOAF Vocabulary Specification...*

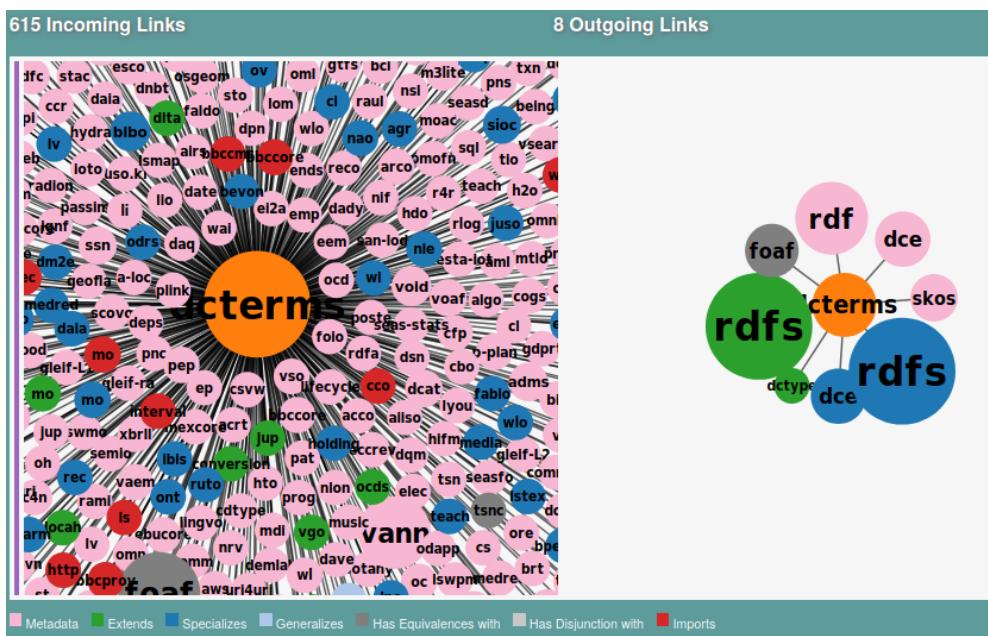


FIGURE 5.2 – DC Terms : représentation des ontologies l’utilisant (à gauche) et de celles qu’elle utilise(à droite) [Source : <https://lov.linkeddata.es/dataset/lov/vocabs/dcterms>]

de groupes et d’organisations, de personnes, mais aussi de réseaux sociaux.

Les ontologies propres aux institutions patrimoniales utilisent ces ontologies de haut niveau. Ainsi, the Bibliographic Ontology (bibo) utilise à la fois les DC Terms et FOAF. CIDOC-CRM³³, contrairement aux autres ontologies institutionnelles, n’est pas de bas niveau, mais de haut niveau. En effet, elle souhaite pouvoir décrire n’importe quel type d’objet : elle dispose de quatre-vingt-cinq classes et de plus de deux cent cinquante propriétés.

Les FRBR et le modèle de données de la BnF permettent de conclure sur l’importance des ontologies dans le Web de données pour les institutions. En effet, son modèle de données étant basé sur les FRBR, tous types de relations sont nécessaires pour relier l’œuvre aux points d’accès — autorités, sujets, dates, ... Ces relations sont exprimées par des ontologies publiques : l’exposition RDF des données permet alors l’utilisation des URIs des ontologies et une description fine du document décrit et de son contexte. Treize ontologies sont ainsi utilisées³⁴, non complètement, mais partiellement, uniquement pour les propriétés nécessaires à la BnF :

- FOAF pour les autorités ;
- RDF pour exprimer les instances — avec `rdf:type` — ;

33. Voir Annexe L : Ontologies de haut et de bas niveaux (Figure L.3 : Une ontologie applicative : C4O)

34. Voir Figure 5.3 : Le modèle de données de la BnF.

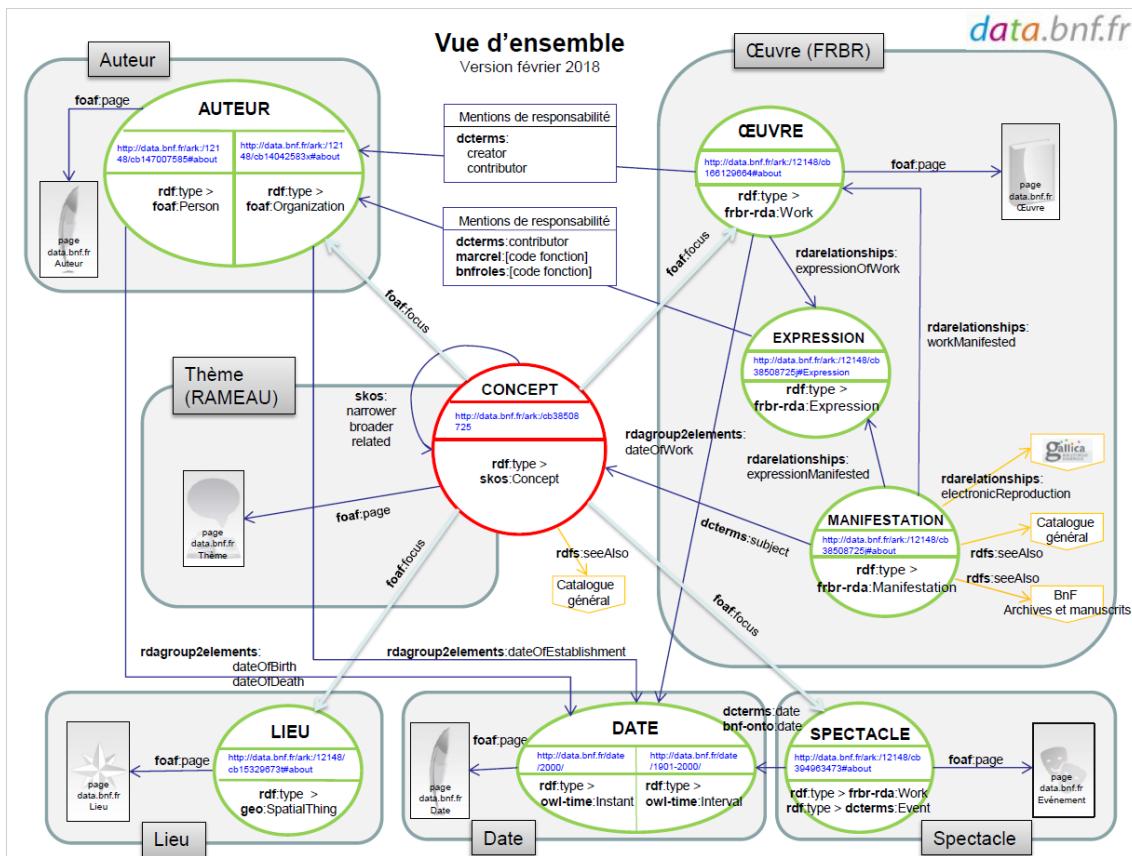


FIGURE 5.3 – Le modèle de données de la BnF [Source : https://data.bnf.fr/images/modèle_donnees_2018_02.PNG]

- RDFS
- SKOS pour créer le concept préférentiel et définir les sujets proches
- DC Terms pour les descriptions bibliographiques simples
- Geo pour les descriptions géographiques
- FRBR-RDA, RDAGroup2elements et RDAreferences pour exprimer les entités du modèle FRBR
- des ontologies internes, BNF-onto et BNFRoles
- l'extension RDF OWL-time
- enfin une ontologie du langage de catalogage MARC MARCrel

L'exemple de la BnF montre qu'il n'y a plus de référentiel unique propre à une institution, mais seulement un ensemble de données dispersées dans le Web de données — utilisation de RAMEAU notamment — avec des ontologies qui permettent la production de sens sur les liens entre les ressources.

Chapitre 6

Relier ses données à Wikidata : l'exemple de l'alignement des personnes physiques de l'INA

Si certaines institutions patrimoniales ont fait le choix d'exposer l'ensemble de leur données sur le Web avec le Web de données, certaines ne font pas ce choix. C'est le cas de l'INA. En effet, l'Institut se concentre sur la mise en cohérence et la centralisation de ses données, tout en les enrichissant le plus possible. Dans ce but d'enrichissement des données, l'INA, nous l'avons évoqué au Chapitre 3 : Les référentiels à l'INA, récupère des données complémentaires auprès de sociétés et créé ainsi du lien avec ces dernières par la conservation dans les bases de données des identifiants propres à chaque document et à chaque société. Cependant, les données fournies sont spécifiques à un domaine d'activité et restreintes. Par conséquent, l'utilisation du Web de données, avec des données ouvertes et accessibles à tous, permet la liaison avec plusieurs jeux de données et la récupération des informations manquantes.

Wikidata est une base de connaissances collaborative RDF concentrant le plus de données et de connaissances sur le monde. L'INA utilise ainsi les identifiants de cette base comme liens vers les données qui y sont stockées¹. Ces identifiants peuvent être ajoutés dès l'opération de catalogage des documents audiovisuels ; ou bien ajoutés *a posteriori* par un alignement.

Pour réaliser cet alignement entre des données institutionnelles et un jeu de données externes du Web sémantique, plusieurs outils existent :

- Des logiciels comme Open Refine² permettent d'aligner des données avec Wiki-

1. Le Chapitre 7 : Les labyrinthes comme réseaux de données et de liens évoque en détail les possibilités offertes par Wikidata dans le parcours des liens et du graphe.

2. Ce logiciel est notamment utilisé par les archives Nationales pour aligner les personnes de la base

data en un clic sur l’interface ; seulement, ce type de logiciel est peu personnalisable dans la constitution des requêtes effectuées.

- Des logiciels ETL permettant de créer des chaînes de récupération et de traitement des données, peu importe le lieu et le type de leur stockage. Talend est ainsi utilisé par l’INA.

De même que l’alignement entre les personnes physiques et le thésaurus de noms communs de l’INA³, celui-ci présente de nombreuses difficultés liées au langage naturel de chaque jeu de données. Plusieurs points de comparaison doivent alors être utilisés, ce qui multiplie les alignements avec Wikidata. De plus, ce type d’alignement révèle de multiples limites, tant techniques que linguistiques.

6.1 Effectuer des requêtes sur Wikidata

Wikidata a son propre modèle de données, mais les données sont représentables et exportables en RDF. Ainsi, Wikidata est totalement intégré dans le Web sémantique. Le langage de requête de RDF, SPARQL Protocol and RDF Query Language (SPARQL), doit permettre un accès rapide aux données par des requêtes parfois complexes. Cependant, d’autres méthodes existent pour accéder aux données, comme une Application Programming Interface (API) fournie par Wikibase via l’URL <https://wikidata.org/w/api.php>.

L’alignement des personnes physiques de l’INA a, pour des raisons expliquées plus bas dans le propos⁴, nécessité l’utilisation de ces deux méthodes afin d’améliorer les délais de récupération des données, et par conséquent d’alignement.

6.1.1 La structure des données de Wikidata

La compréhension de la structure des données de Wikidata, particulière, est nécessaire pour effectuer des requêtes précises et efficaces. Bien que proche de RDF, cette structure comporte quelques particularités⁵. L’affirmation, la structure de base de Wikidata, se compose d’un élément suivi d’une paire propriété-valeur nommée déclaration. Cette affirmation est très proche, sinon identique, au triplet RDF sujet-prédicat-objet. Un élément peut contenir autant de déclarations que nécessaire et est identifié par une URI unique — comme l’est une instance RDF — désignant une page Wikidata commençant par la lettre Q (la lettre P désigne les propriétés).

Léonore de la Légion d’Honneur avec Wikidata.

3. Voir Chapitre 3 : Les référentiels à l’INA.
4. Voir section 6.4 : Comprendre les limites.
5. Voir Annexe J : Repenser la place du référentiel (Figure J.2 : Le modèle de données de Wikidata).

Wikidata peut, en plus de ces données liées semblables à RDF, contenir des informations plus complexes, complexifiant par conséquent la représentation des informations en RDF et les requêtes associées. Ainsi, les déclarations ne sont pas seulement composées de la paire propriété-valeur. Des qualificatifs et des références peuvent y être ajoutés de manière à préciser la paire, eux-mêmes étant sous la forme d'une paire propriété-valeur et constituant un triplet RDF valeur d'une propriété propriété-valeur.

De plus, de même que les ontologies, des axiomes peuvent être introduits pour les propriétés : une propriété peut n'accepter qu'une seule valeur, ou bien une multiplicité de valeurs ; une propriété peut accepter uniquement des éléments identifiés par une URI et non des littéraux, ...

Enfin, Wikidata propose au début de chaque page le label préférentiel de l'élément, une description sommaire, ainsi que des labels alternatifs, pour chaque langue. De manière à mettre ces données dans son modèle de données, les ontologies telles que RDFS, SKOS ou wikibase⁶ sont utilisées et permettent un accès facilité aux données avec RDF.

6.1.2 Le SPARQL-EndPoint

Un premier accès aux données de Wikidata peut se faire par le SPARQL-EndPoint. SPARQL est le langage de requête de données RDF et permet d'effectuer des requêtes complexes en parcourant les triplets. Deux entrées dans ce SPARQL-EndPoint sont possibles :

- Une interface⁷ est disponible pour y écrire les requêtes puis les exécuter. Une prévisualisation des résultats apparaît au bas de l'interface, et il est possible d'exporter les données en plusieurs formats (Comma Separated Values (CSV), JavaScript Object Notation (JSON), RDF, ...).
- Un accès direct aux données est possible en construisant des URLs de requête avec des paramètres : la requête SPARQL est à insérer après <https://query.wikidata.org/sparql> ? ensuite, le paramètre « format » permet la spécification du format de sortie (*&format=json* par exemple)

L'utilisation du SPARQL-EndPoint est adaptée aux requêtes générées humainement. En effet, plus la requête est complexe, plus le temps de réponse est élevé. Il est très fréquent d'obtenir un *timeout* — il peut être du à la complexité de la requête, mais également à la charge des ressources du service de Wikidata ou du réseau Internet —, ce qui empêche l'utilisation de SPARQL dans un processus d'alignement automatique.

6. Cette ontologie est construite à partir de OWL et RDFS.

7. Disponible à l'URL <https://query.wikidata.org/>

Paramètre	Fonction	Particularités
<i>search</i>	Contient la chaîne de caractères à rechercher	Obligatoire
<i>language</i>	Spécifie la langue du texte à rechercher	Obligatoire
<i>limit</i>	Nombre de résultats	

TABLE 6.1 – Paramètres principaux du module *wbsearchentities* de l'API Wikibase [Source : <https://www.wikidata.org/w/api.php?action=help&modules=wbsearchentities>]

S'ajoutent à ces limites d'autres limites imposées par Wikidata⁸ :

- limitation à cinq requêtes simultanées par adresse IP ;
- soixante secondes de traitement toutes les soixante secondes ;
- cette limitation de temps induit des erreurs qui sont limitées à trente par minute ; passée cette limitation à trente erreurs par minute — qui est atteinte rapidement avec un traitement automatique —, l'adresse IP est bloquée

6.1.3 L'API Wikibase

Face à ces nombreuses limites empêchant une utilisation du SPARQL-EndPoint dans l'alignement de données avec Wikidata, l'API permet de meilleurs temps de réponse et l'absence d'erreurs — exceptées celles générées pour les requêtes ne renvoyant aucun résultat⁹. Cette API de Wikibase est uniquement disponible par la méthode GET d'URLs. Elle est représentée par un ensemble de modules qui permettent, sans avoir à écrire de requête SPARQL, de filtrer les résultats de la recherche.

L'accès se fait par l'URL <https://wikidata.org/w/api.php>. À cette URL peuvent être ajoutés des paramètres¹⁰ :

- *action* : permet d'indiquer le module de l'API Wikibase utilisé ;
- *format* : permet de choisir le format de sortie des résultats (JSON ou XML)

Trois modules ont été utilisés pour réaliser cet alignement :

- *wbsearchentities* permet de rechercher la présence d'une chaîne de caractères dans les libellés et les libellés alternatifs des entités de Wikidata ; plusieurs paramètres sont alors disponibles :

8. Consultables dans la documentation : https://www.mediawiki.org/wiki/Wikidata_Query_Service/User_Manual/fr#Limites_des_requ%C3%AAtes

9. L'outil Open Refine utilise cette API et non le SPARQL-EndPoint. Voir le code source du logiciel : <https://github.com/OpenRefine/OpenRefine>.

10. Les paramètres et les modules présentés ci-dessous sont uniquement ceux utilisés dans le traitement avec Talend pour l'alignement des personnes physiques avec Wikidata.

Paramètre	Fonction	Particularités
<i>ids</i>	Liste des entités à récupérer	
<i>props</i>	Type de données à récupérer	Sont utilisés les libellés, les déclarations et les libellés alternatifs
<i>languages</i>	Spécifie la langue du texte à rechercher	Obligatoire

TABLE 6.2 – Paramètres principaux du module *wbgetentities* de l’API Wikibase [Source : <https://www.wikidata.org/w/api.php?action=help&modules=wbgetentities>]

Paramètre	Fonction	Particularités
<i>entity</i>	Identifiant de l’entité	Obligatoire

TABLE 6.3 – Paramètres principaux du module *wbgetclaims* de l’API Wikibase [Source : <https://www.wikidata.org/w/api.php?action=help&modules=wbgetclaims>]

- *wbgetentities* permet d’obtenir tout ou partie des entités de Wikidata
- *wbgetclaims* permet d’obtenir l’ensemble des déclarations de l’entité demandée

La possibilité de lancer plusieurs requêtes HTTP simultanément — en raison de l’absence de vérification de l’adresse IP — et d’y insérer plusieurs entités — jusqu’à cinquante —, quand le paramètre *ids* est possible, rend l’API Wikibase efficace et plus adaptée à un alignement automatique d’un jeu de données avec Wikidata.

Plusieurs points d’accès vers Wikidata sont possibles : l’un peut être utilisé pour des requêtes uniques et complexes, l’autre est utilisé en cas d’automatisation d’un processus. Ainsi, nous le verrons par la suite, un alignement de données avec Wikidata peut nécessiter plusieurs appels à l’API avant de pouvoir choisir l’entité correspondante, en étant plus rapide que les requêtes effectuées avec le SPARQL-EndPoint.

6.2 Aligner des personnes depuis des données contrôlées

L’INA conserve ses données sous deux formes : des données contrôlées, et des données en texte libre¹¹. Les premières permettent une meilleure exploitation et des traitements

11. Voir Annexe C : Les types de données présents dans les bases de données de l’INA et leur rôle (Figure C.1 : Les types de données présents dans les bases de données de l’INA).

facilités, notamment lors d'un alignement avec un référentiel ou un autre jeu de données¹². Cependant, le contrôle du langage naturel se fait différemment selon le contexte de création et d'utilisation des données. Ainsi, deux jeux de données contrôlés peuvent avoir pour un même concept deux graphies et deux normes différentes, tout en respectant de chaque côté un contrôle du langage qui leur est propre.

Bien que les personnes physiques décrites à l'INA respectent des règles de structure, et le soient dans plusieurs attributs d'une table de base de données, elles ne correspondent pas exactement aux entités de Wikidata qui sont créées selon d'autres règles de graphie. Ainsi, quel que soit le jeu de données, il est nécessaire de lui apporter un premier traitement afin de trouver le lien vers un autre jeu de données.

6.2.1 Choix des déclarations des entités de Wikidata

À partir des mêmes personnes physiques de la section 2.3 : Passer du texte libre à un vocabulaire contrôlé : aligner des notes qualité et un thésaurus de noms communs, il est nécessaire de créer du lien avec Wikidata de manière à les enrichir. Plusieurs informations sont ainsi disponibles pour chaque personne : les dates de naissance et de mort, et la note qualité¹³.

Nom	Sexe	Naissance	Décès	Note qualite
Roberts, Howard	Homme	1929-10-02	1992-06-28	Guitariste jazz. Etats Unis

TABLE 6.4 – Informations disponibles pour Howard Roberts à l'INA

L'entité¹⁴ Howard Roberts comporte plusieurs dizaines de déclarations — paire propriété-valeur. Cependant, la présence de ces déclarations n'est pas uniforme selon les entités de personnes. De manière à aligner les entités Wikidata avec un jeu de données, il est nécessaire de chercher les propriétés les plus communes et le plus souvent présentes, qui puissent correspondre à chacune des informations disponibles dans le jeu de données à aligner.

Ce choix, pour les dates de naissance ou de mort, le sexe ou le pays de citoyenneté, est aisé et correspond à des propriétés utilisées presque systématiquement dans les entités de type personne :

- le sexe de la personne est défini par la propriété P21¹⁵ de Wikidata ; une entité correspond à la valeur

12. L'alignement des notes qualité de l'INA avec un thésaurus interne a déjà démontré la difficulté de l'utilisation du texte libre.

13. Voir Table 6.4 : Informations disponibles pour Howard Roberts à l'INA.

14. Howard Roberts : <https://www.wikidata.org/wiki/Q1631895>

15. Sexe (P21) : <https://www.wikidata.org/wiki/Property:P21>

- la date de naissance est la propriété P569¹⁶; la valeur de cette déclaration est un littéral
- la date de décès est la propriété P570¹⁷; la valeur de cette déclaration est un littéral
- enfin, le pays de citoyenneté est la propriété P27¹⁸; la valeur est une entité de pays

6.2.2 Adapter les données contrôlées pour les valeurs des déclarations

Bien que contrôlés, les jeux de données diffèrent dans la graphie de leurs valeurs¹⁹. Il est par conséquent nécessaire d'effectuer un premier traitement sur les données de l'INA de manière à faciliter leur alignement avec Wikidata.

	Nom	Sexe	Naissance	Décès
INA	Roberts, Howard	Homme	1929-10-02	1992-06-28
Wikidata	Howard Roberts	masculin(Q6581097)	2 octobre 1929	28 juin 1992
		Note qualité	Pays	
		Guitariste jazz. Etats Unis	États-Unis (Q30)	

TABLE 6.5 – Comparaison des informations disponibles pour Howard Roberts à l'INA et sur Wikidata

D'abord, une inversion des noms et prénoms des personnes physiques de l'INA doit être effectuée afin de correspondre au libellé de Wikidata. Il existe des propriétés nom de famille (P734)²⁰ et prénom (P735)²¹: ce formalisme trop important par rapport aux données de l'INA n'aurait pas permis de comparer également les autres libellés qui ne sont pas divisés ainsi. L'inversion du nom et du prénom a par conséquent été réalisée pour pouvoir correspondre au libellé préférentiel de l'entité, ainsi qu'à ses libellés alternatifs.

Ensuite, le pays doit être extrait de la note qualité selon la méthode expliquée à la section 2.3 : Passer du texte libre à un vocabulaire contrôlé : aligner des notes qualité et un thésaurus de noms communs. En cas de pays multiples dans cette note, la ligne et ses informations est répétée autant de fois nécessaire avec un pays par ligne : cela permet de traiter la double nationalité ou les différents pays d'exercice des fonctions de chaque

16. Date de naissance (P569) : <https://www.wikidata.org/wiki/Property:P569>

17. Date de décès (P570) : <https://www.wikidata.org/wiki/Property:P570>

18. Pays de citoyenneté (P27) : <https://www.wikidata.org/wiki/Property:P27>

19. Voir Table 6.5 : Comparaison des informations disponibles pour Howard Roberts à l'INA et sur Wikidata.

20. Nom de famille (P734) : <https://www.wikidata.org/wiki/Property:P734>

21. Prénom (P735) : <https://www.wikidata.org/wiki/Property:P735>

personne. La propriété P27 n’indiquant que le pays de citoyenneté, conserver les multiples pays de la note qualité permet d’augmenter les chances d’alignement avec Wikidata.

Enfin, afin d’éviter l’écueil des différentes notations des dates — en plein texte, avec des tirets ou des barres obliques, avec les horaires, ... —, les dates de naissance et de décès sont réduites à l’année à la fois pour les données de l’INA que pour celles de Wikidata. Il est en effet possible de considérer qu’une personne et une entité partageant le même libellé et ayant les mêmes années de naissance et de décès sont équivalentes.

	Nom	Sexe	Naissance	Décès
INA	Howard Roberts	Homme	1929	1992
Wikidata	Howard Roberts	masculin(Q6581097)	1929	1992
	Note qualité	Pays		
INA	Guitariste jazz. Etats Unis	Etats-Unis		
Wikidata		États-Unis (Q30)		

TABLE 6.6 – Comparaison des informations disponibles pour Howard Roberts à l’INA et sur Wikidata après un premier traitement

Quand les deux jeux de données présentent la même forme et la même graphie pour des propriétés équivalentes, l’alignement peut débuter²². Cependant, les personnes physiques de l’INA peuvent ne pas avoir de date de naissance connue, un pays absent, ... Afin de ne pas comparer ces données manquantes avec celles certainement présentes de Wikidata, il convient de regrouper les personnes physiques selon les caractéristiques qui sont remplies, pour adapter les requêtes et les comparaisons.

6.2.3 Effectuer l’alignement par de multiples requêtes

Le premier traitement effectué ci-dessus ne suffit cependant pas à rendre les données alignables entre-elles. En effet, *Q6581097* et *Homme* ne sont pas similaires, tout comme le pays avec l’identifiant Wikidata. Il est par conséquent nécessaire d’attribuer ces identifiants Wikidata dans le jeu de données de l’INA. Ainsi, les mentions de genre, « Homme » et « Femme » sont, quand elles sont présentes, remplacées par l’identifiant des entités « masculin »²³ et « féminin »²⁴ de Wikidata.

L’ajout de l’identifiant du pays est également nécessaire. Ces pays se comptant par centaines, l’alignement doit être réalisé automatiquement à partir d’une requête dans le SPARQL-EndPoint. En effet, une seule requête, non répétée ensuite, est nécessaire

22. Voir Table 6.6 : Comparaison des informations disponibles pour Howard Roberts à l’INA et sur Wikidata après un premier traitement.

23. masculin (Q6581097) : <https://www.wikidata.org/wiki/Q6581097>

24. féminin (Q1775415) : <https://www.wikidata.org/wiki/Q1775415>

pour récupérer l'ensemble des pays de Wikidata, leur identifiant, leur libellé préférentiel et leurs libellés alternatifs²⁵. Les pays récupérés, ainsi que les pays de l'INA, voient leur ponctuation et leur accentuation supprimées, et les majuscules transformées en minuscules afin d'arriver à un rapprochement le plus grand possible entre les graphies d'un même pays. Les États-Unis indiqués dans les données de l'INA sont alors alignés avec le même identifiant que le pays de citoyenneté de l'entité Howard ROBERTS de Wikidata.

```

SELECT DISTINCT ?country ?countryLabel ?altLabel
WHERE
{
?country wdt:P31 wd:Q3624078 .
#en excluant les États historiques n'existant plus
FILTER NOT EXISTS {?country wdt:P31 wd:Q3024240}
#ainsi que les anciennes civilisations
FILTER NOT EXISTS {?country wdt:P31 wd:Q28171280}
OPTIONAL{?country skos:altLabel ?altLabel . FILTER (lang(?altLabel) = "fr")} .

SERVICE wikibase:label { bd:serviceParam wikibase:language "fr" }
}
ORDER BY ?countryLabel

```

FIGURE 6.1 – Requête SPARQL de récupération des pays

À la suite de ces deux précédents alignements, l'alignement de la personne elle-même devient possible : les quatre points de comparaison sont désormais potentiellement comparables. La rédaction d'une requête SPARQL par personne, avec des filtres selon les données disponibles à la comparaison, demande beaucoup de temps de traitement et de recherche. Pour cette raison, l'API Wikibase est utilisée en deux étapes²⁶ :

- l'utilisation du module *wbsearchentities* permet la recherche par le nom de la personne²⁷ ; l'ensemble des identifiants d'entités retournés sont alors stockés pour l'étape suivante
- il est ensuite possible, avec le module *wbgetclaims*²⁸, d'obtenir les déclarations de chaque entité, et par conséquent les valeurs des propriétés recherchées — P21, P27, P569 et P570

La récupération des valeurs de ces propriétés permet l'obtention de la Table 6.6 : Comparaison des informations disponibles pour Howard Roberts à l'INA et sur Wikidata après un premier traitement puis l'alignement de la personne physique de l'INA avec la bonne

25. Voir Figure 6.1 : Requête SPARQL de récupération des pays.

26. Il a été constaté qu'avec le même logiciel — Talend — et la même machine, l'exécution de deux requêtes avec l'API prend quarante fois moins de temps qu'une unique requête dans le SPARQL-EndPoint.

27. Pour Howard ROBERTS, cette recherche est la suivante : <https://www.wikidata.org/w/api.php?action=wbsearchentities&language=fr&search=howard%20roberts&format=json>.

28. Voir <https://www.wikidata.org/w/api.php?action=wbgetclaims&entity=Q1631895&format=json>.

entité de Wikidata, si cette dernière existe. Cependant, la seule utilisation de ces quatre points de comparaison permet qu'un petit nombre d'alignements. En effet, les données de l'INA comportent peu ou pas de dates de naissance et de décès : la réalisation d'un alignement sur les seuls genre et pays de citoyenneté ne peut pas être satisfaisante et ne permet pas une bonne gestion des homonymes.

La présence de données contrôlées dans une institution ne permet pas à elle seule de permettre un alignement avec des référentiels ou des jeux de données externes : bien que structurées et contrôlées, les données diffèrent d'une institution à une autre, d'un jeu de données à un autre, par leur graphie, leur norme de rédaction, ... Il sera par conséquent toujours nécessaire d'effectuer des traitements préalables de données que l'on souhaite mettre en relation avec d'autres.

6.3 Aligner des personnes depuis du texte libre

L'utilisation de données contrôlées pour effectuer un alignement n'est pas suffisante ; les notes en texte libre doivent alors être utilisées car elles comportent des informations permettant de réaliser cet alignement. Cependant, l'utilisation d'un texte libre nécessite plus de traitement et d'adaptations. Ainsi, le seul alignement de la fonction extraite des notes qualité avec son équivalent dans Wikidata est confronté aux classes et sous-classes de Wikidata, et une nouvelle fois aux différents contrôles et règles s'appliquant à chaque jeu de données. De cet alignement dépend l'alignement des personnes.

6.3.1 Aligner les fonctions avec Wikidata

Les métiers et fonctions de personnes sont réunis, sur Wikidata, sous l'entité des *Activités humaines* (Q61788060)²⁹. L'utilisation de cette classe, très générique, est nécessaire afin d'obtenir l'ensemble des fonctions et des métiers qui existent dans Wikidata. Seulement, les entités dépendant directement ou indirectement de cette entité sont très nombreuses³⁰ : un tri de ces entités n'est pas possible car l'étendue des fonctions des notes qualité n'est pas connue précisément.

L'extraction de l'ensemble des entités de Q61788060 doit permettre d'obtenir l'identifiant de chaque entité, ainsi que le libellé préférentiel et les libellés alternatifs. Une telle requête dans le SPARQL-EndPoint n'est pas possible car elle est trop lourde. Ainsi, la scission de la requête en plusieurs étapes est nécessaire :

29. Activités humaines (Q61788060) : <https://www.wikidata.org/wiki/Q61788060>

30. La requête SPARQL de comptage de ces entités indique plus de 1100000 entités dépendant des *Activités humaines*.

- d'abord, il y a récupération, par la propriété P31, des entités dépendant directement des *Activités humaines*
- ensuite, la récupération des entités étant des sous-classes (propriété P279) de Q61788060 a lieu ;
- ce n'est qu'à partir de ce niveau de hiérarchie que la récupération des entités liées à ces entités sous-classes d'*Activités humaines* peut avoir lieu : pour chacune de ces entités récupérées dans la seconde étape, une requête SPARQL est effectuée pour obtenir les identifiants et les libellés des entités liées. Ainsi, pour obtenir les entités liées à l'entité *Profession* (Q28640)³¹, la requête est la suivante :

```
select ?f ?fLabel ?altLabel
where
{
?f wdt:P31/wdt:P279* wd:Q28640.
OPTIONAL{?f skos:altLabel ?altLabel FILTER (LANG(?altLabel) = "fr")}
SERVICE wikibase:label { bd:serviceParam wikibase:language "fr". }
}
```

Le modèle de données de Wikidata étant sous la forme d'un graphe, les relations entre les entités sont multiples et très nombreuses. Pour saisir le maximum de fonctions et de métiers, il est donc nécessaire de trouver l'entité la plus haute dans la hiérarchie, celle qui comporte le moins de liens entrants, qui est ici Q61788060 pour les *Activités humaines*³².

Cependant, de même que dans la section 6.2 : Aligner des personnes depuis des données contrôlées, les jeux de données n'ont pas les mêmes règles concernant les graphies des noms. Alors, l'application de règles identiques doit être faite pour chaque jeu de données, celui de l'INA, et celui des entités extraites de Wikidata : suppression des majuscules, de la ponctuation et de l'accentuation, transformation des féminins en masculins, et des pluriels en singuliers. Cette normalisation identique pour chaque jeu de données permet un rapprochement des deux graphies, et une amélioration de l'alignement qui est réalisé ensuite entre la fonction de la note qualité et le libellé de l'entité Wikidata.

6.3.2 Utiliser la hiérarchie de Wikidata pour aligner les personnes

L'alignement de la fonction avec son entité équivalente de Wikidata a permis de créer le point de comparaison utile à l'alignement des personnes. En effet, chaque entité de type *Humain* de Wikidata peut avoir une propriété *Occupation* (P106)³³. Cette propriété peut accepter plusieurs valeurs qui sont obligatoirement des entités Wikidata et

31. Profession (Q28640) : <https://www.wikidata.org/wiki/Q28640>

32. Voir Figure 6.2 : Modélisation des entités Wikidata dont *Humoriste* est issu.

33. Occupation (P106) : <https://www.wikidata.org/wiki/Property:P106>

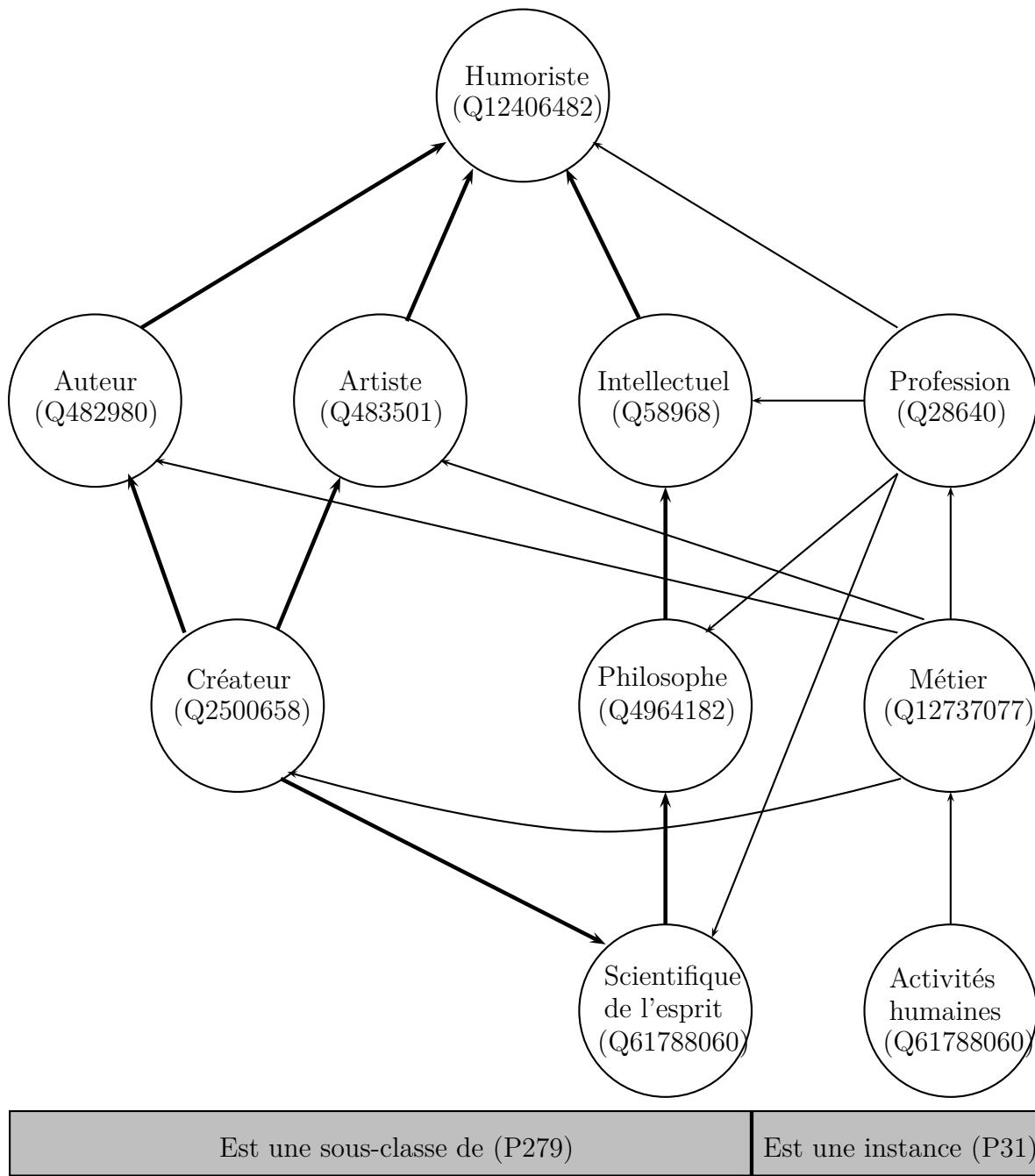


FIGURE 6.2 – Modélisation des entités Wikidata dont *Humoriste* est issu

le plus souvent des entités instances ou sous-instances de *Activités humaines*. Ainsi, la comparaison entre la personne de l'INA et celle de Wikidata peut s'effectuer grâce à cette propriété.

Cependant, il est fréquent de trouver des entités avec des valeurs de P106 n'étant pas celle de la fonction de la note qualité. Cela est dû à plusieurs facteurs :

- la valeur entrée dans Wikidata est trop spécifique : par exemple, la valeur peut être *Chanteur pop*, alors qu'il est simplement indiqué *Chanteur* à l'INA et que

c'est cette entité *Chanteur* qui a été aligné précédemment

- au contraire, la valeur de Wikidata peut être plus générique que celle de l'INA

De manière à pouvoir aligner l'entité et la personne qui correspondent, malgré cette différence de hiérarchie de la fonction dans Wikidata, il est nécessaire d'utiliser la classe dont est issue la valeur de P106, ainsi que les sous-classes de cette même valeur. Cette méthode permet de générer plus de possibilités de comparaison, sans créer de généralisation trop forte au niveau supérieur qui induirait des erreurs d'alignements. Pour le cas d'*Humoriste*, la personne de l'INA sera comparée depuis sa fonction avec :

- *Humoriste* (Q12406482)
- les entités dont il est l'instance :
 - *Intellectuel* (Q58968)
 - *Artiste* (Q483501)
 - et *Auteur* (Q482980)
- les instances dont il est la classe :
 - *Bouffon* (Q215548)
 - *Dessinateur humoristique* (Q1114448)
 - et *Bouffon* (Q15037346)

Avec du texte libre, un alignement devient plus compliqué : d'une part, la graphie diffère selon les jeux de données et le langage naturel utilisé ; d'autre part, le niveau de précision de la description est variable — ici le niveau de précision de la fonction d'une personne. Le périmètre de l'alignement devient plus large au risque d'introduire des erreurs d'alignement.

Au terme de l'alignement réalisé à partir des données contrôlées, puis de celui effectué à partir des notes qualités, toutes les personnes qui ont une entité Wikidata n'ont pas pu être alignées. Un dernier alignement sur le seul état civil aurait pu être envisagé, mais il aurait certainement causé beaucoup d'erreurs en raison des homonymes.

6.4 Comprendre les limites

Le lien offert par l'identifiant Wikidata est précieux grâce aux nombreuses informations et données qu'il peut fournir. Seulement, pour obtenir ce lien automatiquement, un alignement doit être réalisé grâce à des comparaisons de données qui se complexifient en fonction de la nature des données, de leur structure et surtout de leur similarité de part et d'autre de l'alignement.

Ces limites, associées à celles du Web sémantique et de Wikidata, réduisent l'efficacité de cet alignement : l'INA possède environ 330000 personnes physiques qui n'ont pas

d'identifiant Wikidata associé ; l'enjeu de cet alignement est donc important. Cependant, seuls 31000 identifiants de Wikidata ont pu être récupérés : 90% des personnes physiques n'ont par conséquent pas pu être alignées.

6.4.1 Les raisons de l'absence d'alignement

Les raisons de l'absence d'alignement des personnes physiques de l'INA avec Wikidata sont multiples et diverses, et principalement liées à la donnée elle-même, dans sa forme, sa graphie ou sa structure.

D'abord, de nombreuses personnes physiques de l'INA ont des données éparses, sans dates de naissance ou de décès, sans genre. Quand la fonction indiquée n'est pas réellement une fonction, mais une notion d'appartenance à un événement ou à une famille, cette personne se retrouve sans point de comparaison possible. Ainsi, Abdesslam GUE-ROUAZ ne dispose ni de dates ni de fonction utilisable dans l'alignement — la note qualité est « Attentat, Maroc 1994.Maroc ». Cette appartenance à un événement n'est pas une valeur correspondant à la propriété *Occupation* (P106) servant aux comparaisons sur les fonctions des notes qualité.

Ensuite, la divergence des états-civil est une autre source d'empêchement des alignements. Si les noms et prénoms français — qui sont par conséquent ni traduits à l'INA ni sur Wikidata — ne sont pas massivement concernés, ceux étrangers le sont car les traductions varient selon le traducteur et les normes de catalogage et de remplissage. C'est pourquoi Tomasz POPAKUL du jeu de données de l'INA, réalisateur selon la note qualité, ne peut pas être aligné avec son entité Wikidata Q19269951. En effet, le libellé de Wikidata est Tomek POPAKUL. Malgré la concordance de la fonction de la note qualité avec la valeur de la propriété P106, cette personne n'a pas pu être alignée en raison de différences de traduction de l'état civil.

L'alignement des fonctions est également une des raisons de l'échec des alignements de personnes. En effet, comme constaté au Chapitre 2 : L'arbre, un vocabulaire contrôlé hiérarchique, les premiers termes des notes qualité peuvent être génériques et bloquer les traitements. Ainsi, Cameron ALBORZIAN, « Ancien mannequin, maître de yoga » dans la note qualité, ne peut pas être aligné avec son entité Wikidata Q5026166 en raison de l'échec de l'alignement de la fonction. La propriété P106 de Q5026166 comporte la valeur *Mannequin* (Q4610556) alors que la fonction de l'INA est *Ancien mannequin*.

Cependant, la majorité des personnes physiques non alignées avec Wikidata est due à l'absence de ces personnes dans Wikidata : l'INA documente chaque personne de

générique, du scripte au cadreur, réalisateur, ... Ces personnes, peu ou pas connues, ne font par conséquent pas l'objet d'une entité sur Wikidata.

6.4.2 Les limites du Web sémantique

Les limites imposées par le Web sémantique compliquent également le processus d'alignement entre deux jeux de données. Les performances de SPARQL, et plus généralement de RDF, sont difficilement compatibles avec des requêtes de masse et complexes. Les temps de réponse, dès lors qu'un appel au service de langage de Wikibase est effectué, que trop de clauses optionnelles ou obligatoires sont introduites, ou bien que le nombre de résultats est trop important, augmentent considérablement. Ainsi, pour éviter ces temps de réponses trop importants, Wikidata a introduit une limite à soixante secondes. La présence de cette limite constraint à ne pas utiliser le SPARQL-EndPoint dans le cadre d'un processus automatique d'alignement de données, et à se tourner vers une autre solution, l'API Wikibase qui offre des performances supérieures, sans retour d'erreurs liées aux temps de réponses.

La rapidité d'un service sur le Web est essentielle et fait partie, comme le rappelle Gautier POUPEAU dans son billet de blog³⁴, des axes indispensables du Big Data³⁵. Ainsi, la cinquième étoile de la classification de Tim BERNERS-LEE, composée de RDF, paraît difficilement atteignable et utilisable universellement : l'API Wikibase renvoie les résultats sous forme de JSON ou de XML, et non en RDF. Le symptôme des difficultés de RDF de s'imposer avec son langage de requête SPARQL est son utilisation et sa maîtrise par peu de développeurs³⁶ ; cependant, le modèle en graphe³⁷ développé à partir de la réflexion de Tim BERNERS-LEE n'est pas remis en cause et est de plus en plus utilisé.

Enfin, le Web sémantique, et Wikidata, se placent dans une vision d'interopérabilité à partir des données dont nous avons développé les avantages avec la disparition de la notion de référentiel. Cette intéropérabilité, théoriquement, doit permettre de s'affranchir des règles du langage naturel et de proposer des données partageables et réutilisables par tous. Cependant, ces données devront toujours subir des traitements et des transformations pour être intégrées dans des systèmes documentaires ou des jeux de données. La

34. G. Poupeau, *Au-delà des limites, que reste-t-il concrètement du Web sémantique ?*, Les Petites Cases, 6 oct. 2018, URL : <http://www.lespetitescases.net/au-delà-des-limites-que-reste-t-il-concrettement-du-web-semantique> (visité le 01/08/2020).

35. « Parmi les trois axes qui définissent traditionnellement le Big Data, vitesse, volume et variété (les « 3V »), les deux premières caractéristiques ne sont pas encore atteintes par ces technologies » in *Ibid.*

36. « il est à craindre que les technologies du Web sémantique restent des technologies de niche maîtrisées par peu de développeurs » in *Ibid.*

37. Voir Partie III : CENTRALISER. Le référentiel, clé de voûte et pivot (depuis le milieu des années 2010).

normalisation des libellés et des dates des entités de Wikidata en est un exemple, de manière à pouvoir être mis en concordance avec d'autres données, celles de l'INA.

Face aux possibilités d'obtenir un enrichissement de ses propres données à partir de jeux de données externes et publics comme Wikidata, les institutions patrimoniales sont de plus en plus nombreuses à utiliser le Web de données. Mais, au-delà de ces avantages, les limites et les difficultés sont réelles : faut-il alors bâtir un modèle de données, ou un référentiel, internes sur un modèle du Web de données ? Est-il possible d'utiliser un jeu de données externe comme référentiel interne, adapté aux besoins et aux usages spécifiques à une institution ?

Le traitement de la donnée sera toujours nécessaire pour pouvoir faire correspondre cette donnée aux usages de chacun selon des besoins spécifiques. Bien que Wikidata ou d'autres jeux de données du Linked Open Data soient complets et se veulent universels dans leur domaine, les besoins et les usages sont tous différents ; ainsi, les normes et les conventions de rédaction des données varient selon ces derniers dans chaque institution, rendant impossible l'écriture du dictionnaire universel dans lequel chercher les données nécessaires.

Cette seconde partie a mis en évidence une importante évolution dans la place des référentiels et les usages qui en sont faits. Ils ne se trouvent plus en marge des systèmes documentaires : les efforts donnés à les partager pour les réutiliser montre un glissement vers le centre de ces systèmes, sans toutefois l'atteindre. Le lien, plus que la donnée elle-même, est devenu un enjeu pour toutes les institutions : un enrichissement de ses propres données est possible. Ces partages constants ont montré l'importance des formats et des protocoles d'échanges communs et utilisés par tous.

Deux cas de figure généraux de la place des référentiels dans les institutions et des liens qu'ils entretiennent entre eux sont apparus : d'une part il peut y avoir une simple réutilisation des données d'une institution dans une autre (Figure 6.3 : Réutilisation d'un référentiel entre deux institutions) ; d'autre part, le Web sémantique a permis la création de nouveaux référentiels avec lesquels il est possible de créer du lien (Figure 6.4 : Utilisation commune d'un jeu de données du Web de données).

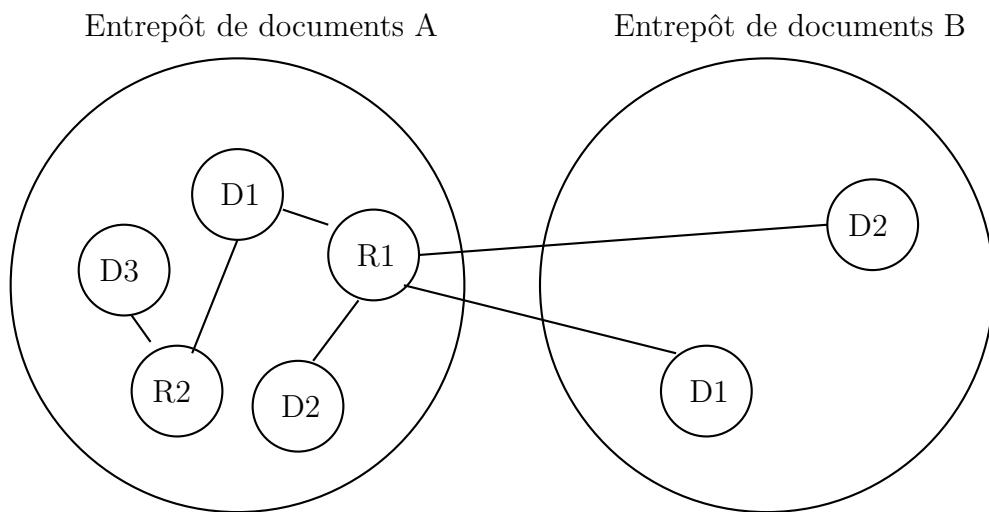


FIGURE 6.3 – Réutilisation d'un référentiel entre deux institutions (R : Référentiel ; D : Document)

Cependant, l'unique utilisation d'un seul référentiel ne permet pas à une institution de combler l'ensemble de ses besoins et de ses usages : plusieurs liens doivent alors être établis avec divers jeux de données, ce qui peut être réalisé par les ontologies et le passage de lien en lien. Bien que le lien soit devenu un élément structurant du Web sémantique, il devient d'autant plus précieux quand il se trouve lui-même objet d'une grande entité³⁸.

38. Comme il a été montré sans autre précision avec VIAF dans le Chapitre 4 : Le web de données : une exposition commune des référentiels, et ce qui fera la suite de notre propos en Partie III : CENTRALISER. Le référentiel, clé de voûte et pivot (depuis le milieu des années 2010).

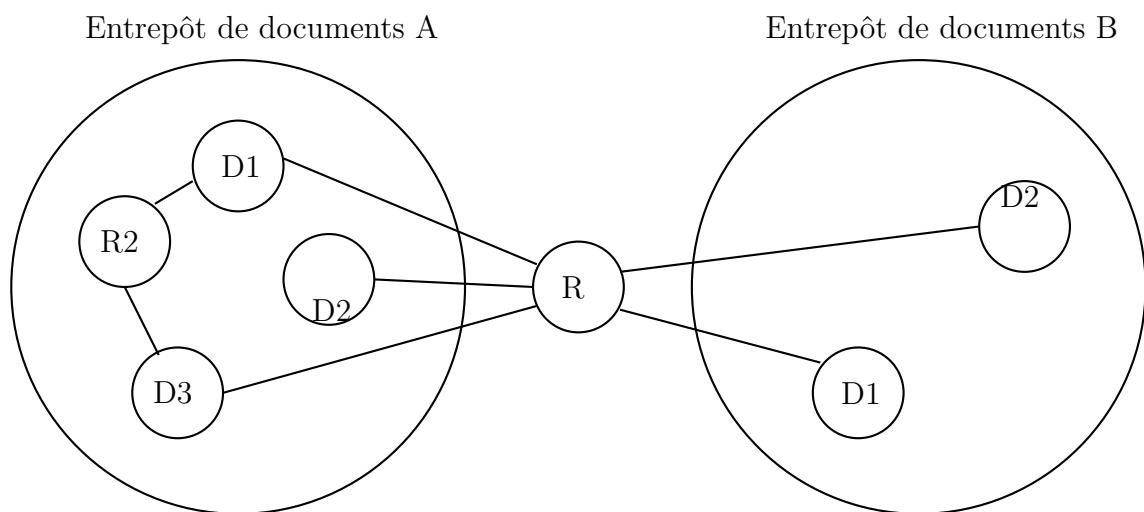


FIGURE 6.4 – Utilisation commune d'un jeu de données du Web de données (R : Référentiel ; D : Document)

Troisième partie

CENTRALISER. Le référentiel, clé de voûte et pivot (depuis le milieu des années 2010)

Le Web de données a initié une pratique nouvelle dans la conception des modèles de données. Le lien devient une ressource indispensable qui permet aux institutions et aux systèmes documentaires de déstructurer leurs jeux de données afin de les enrichir, de les partager, de mieux les valoriser. Alors, la création ou l'utilisation d'un référentiel devient quasiment inutile — sauf en cas de besoins très spécifiques comme c'est le cas à l'INA. Le parcours des liens, au travers un graphe, permet bien plus que ce que permet un référentiel interne stocké dans une base de données relationnelles.

Le modèle de données comme réseau de liens et de noeuds n'est pas une notion récente, mais elle a trouvé son application et son utilité avec l'apparition de l'informatique et du Web. Ce réseau de noeuds et de liens est alors, depuis la seconde moitié des années 2010, de plus en plus adopté lors des refontes de systèmes d'information, afin de s'adapter aux nouvelles pratiques des utilisateurs sur le Web et aux nouveaux besoins qui en découlent. L'application de ce modèle-réseau en interne créé alors du Linked Enterprise Data (LED), l'équivalent du Linked Open Data (LOD) pour le Web de données.

Le *Lac de données* mis en œuvre depuis 2015 à l'INA est un LED qui reprend les principes de déconstruction du document et de l'information au profit de la création de grandes entités et de multiples liens, ce qui permet une grande adaptation aux besoins actuels et futurs auxquels les données doivent ou devront répondre. Les systèmes documentaires ne sont ainsi plus pensés à partir des besoins, mais des données. Le référentiel dans le *Lac de données* prend une place centrale puisqu'il permet la description de l'ensemble des instances et est essentiel aux nouveaux besoins relatifs à l'intelligence artificielle. Cependant, la fusion de multiples référentiels au sein d'un système uniformisé et centralisé est une tâche complexe qui doit éviter les doublons et effacer les différences de structure et de graphie qui existaient auparavant.

Chapitre 7

Les labyrinthes comme réseaux de données et de liens

La multiplication des liens et de ceux possibles dans le Web de données entraîne une désorganisation — aux yeux d'un humain — des informations et des référentiels dans ce Web de données. Les chemins à emprunter deviennent multiples et provoquent une ivresse de rebonds et d'informations chez l'utilisateur. Les interfaces de visualisation structurent l'ensemble des informations et des liens du Web de données qui est devenu un Web où seule une machine peut se repérer rapidement et naviguer aisément. Du modèle du graphe d'un jeu de données, le Web de données a permis de créer un graphe à l'échelle du Web, accessible par tous et en tous points, depuis n'importe lequel des jeux de données, des référentiels ou des institutions.

La notion de graphe, de réseau de données, découle des nombreuses théories d'arbres de classifications et de descriptions des précédents millénaires. La constatation des limites et de l'échec de ces arbres a conduit à la théorisation, puis l'adoption dans le Web de données et par le milieu bibliothéconomique d'abord, du labyrinthe et du modèle-réseau de données. Le lien devenant l'essence-même de ces réseaux de données, de nouveaux types de référentiels ont vu le jour, notamment les hubs de liens qui centralisent les liens et quelques données d'autorités autour d'un même identifiant. Wikidata, d'abord réceptacle structuré des données et des informations des Wikipédias, devient rapidement le hub de liens et d'identifiants le plus utilisé.

7.1 Du modèle encyclopédique aux graphes de données

Au Moyen-Âge, la dogmatique de l'arbre porphyrien domine. Ce n'est qu'à la Renaissance que le savoir est conçu comme ouvert. L'arbre était pensé selon le monde qui

était lui-même pensé comme un cosmos clos, ordonnée ; ce même arbre était par ailleurs pensé comme une finitude inaltérable de sphères. Cependant, la pensée de Copernic influe la façon de concevoir le savoir : ce dernier s'efforce de mimer le système planétaire avec ses perspectives variables, des orbites qui deviennent des ellipses, ...

L'encyclopédie n'est alors plus un amas de connaissances réelles et légendaires, elle devient un index devant décrire le monde et les connaissances, le classifier. La tension pesant sur ce modèle encyclopédique et la quantité infinie de connaissances conduit à son éclatement au profit d'une forêt où tout est ou peut être relié selon les choix du lecteur.

7.1.1 Vers les labyrinthes (Renaissance)

L'évolution majeure de la Renaissance, faisant suite aux arbres porphyriens puis lulliens, est la nouvelle conception de la structure des éléments du monde : avec Porphyre et ses successeurs, seuls les accidents et les accidents sont classifiés ; avec la Renaissance, de multiples index d'encyclopédies naissent, accompagnés de réflexion sur les manières d'ordonner le savoir¹. Toute la Renaissance va se concentrer sur cette classification du savoir.

La première grande encyclopédie tentant cette classification du savoir est l'*Encyclopaedia septem tomis distincta* de Johann ALSTED en 1620² : l'index devient la substance-même de cette œuvre. Cette encyclopédie s'inscrit dans la période pansophique de la Renaissance, dans laquelle la réflexion sur une sapience universelle qui aurait toute l'étendue du savoir, est vive. Si l'arbre de Porphyre voulait simplement être un dictionnaire, un moyen de définir la science, l'index pansophique inspire lui à classifier cette science, et s'éloigne par cela du dictionnaire. Cette période pansophique marque bien l'arrêt de la conception hiérarchique du savoir conçue comme moyen de définir : un autre moyen de décrire ce savoir est possible, en étant plus efficace.

Ce nouveau moyen part de la constatation que de multiples chemins peuvent mener à un même savoir. Francis BACON le constate dès 1620 dans l'*Instauratio Magna*³ puis en 1626 dans le *Sylva sylvarum*. Il n'est alors plus question d'arbre unique, mais d'arbres multiples, de labyrinthe avec des chemins ambigus, des ressemblances trompeuses, des spirales et des noeuds complexes⁴. La forêt est un amas de sujets, on n'y trouve plus,

1. « Nous n'avons plus affaire à une classification de substances et d'accidents, mais à l'index d'une encyclopédie possible et à la tentative de proposer un ordonnancement du savoir » in U. Eco, *De l'arbre au labyrinthe...*

2. Johann Heinrich Alsted, *Encyclopaedia septem tomis distincta...* 2 t., Herbornae Nassoviorum, 1630.

3. Francis Bacon, *Instauratio magna*, Londini, 1620.

4. « obliquae et implexae naturarum spirae et nodi » in *Ibid.*

mais on découvre de nouvelles relations, ce que l'on ne sait pas encore et ce que l'on cherchait pas. Cependant, la « tension entre l'arbre et le labyrinthe »⁵ ne faiblit pas : John WILKINS, à la fin des années 1660, est mis en échec devant ses classifications du savoir qui ne parviennent pas à classer les sujets ; une table d'index immense est alors créée pour résoudre cette difficulté.

La masse des connaissances à classer étant immense, l'encyclopédie devient un inventaire général des connaissances, incapable de toutes les saisir : Leibniz comprend bien que l'entreprise encyclopédique peut être infinie en raison du nombre de renvois à créer pour s'adapter aux perspectives infinies d'accès à une connaissance. Le labyrinthe prend alors tout son sens : une connaissance est accessible par de multiples points d'accès et ne fait pas partie d'une hiérarchie stricte. Dans le « Discours préliminaire » de l'Encyclopédie⁶, l'arbre porphyrien et la pensée artistotélicienne sont totalement remis en question.

« Le système général des sciences et des arts est une espèce de labyrinthe, de chemin tortueux, où l'esprit s'engage sans trop connaître la route qu'il doit tenir.⁷ »

L'encyclopédie totale et universelle ne pourra jamais voir le jour en raison de son ampleur ; elle n'est qu'utopie de la connaissance. Cette utopie, toujours visible aujourd'hui — les projets Wikidata ou Wikipédia se veulent le reflet de notre monde —, ne cessera pas en raison du caractère culturel qu'elle contient : plus que le reflet de notre monde, elle est le reflet de notre culture et de nos cultures spécifiques. L'encyclopédie sert cependant à créer des portions d'encyclopédie, en vue de réaliser des classifications spécifiques à un domaine.

7.1.2 Des labyrinthes aux graphes

L'arrivée de la pensée classificatoire au labyrinthe a eu lieu à plusieurs reprises, à chaque échec du modèle de l'arbre, notamment avec Porphyre où chaque pas dans l'arbre régénérat sans cesse l'arbre des différences : il n'y a par conséquent, pas un, mais un nombre infini d'arbres selon leur contexte. Il existe plusieurs types de labyrinthes dont les trois principaux sont présentés par Umberto ECO :

- Le plus ancien est un labyrinthe classique unicursal, dit de Knossos (Figure M.1 : Le labyrinthe de Knossos) : la seule possibilité est d'atteindre son centre ; en raison de cette caractéristique, il ne peut pas se rapporter à un modèle encyclo-

5. U. Eco, *De l'arbre au labyrinthe...*

6. *Encyclopédie, ou Dictionnaire raisonné des sciences, des arts et des métiers*, par une société de gens de lettres. Mis en ordre & publié par M. Diderot..., & quant à la partie mathématique, par M. d'Alembert... Dir. Denis Diderot et Jean Le Rond D'Alembert, avec la coll. d'Antoine-Claude Briasson, et al., 35 t., 1751.

7. *Ibid.*, Discours préliminaire.

pédique ni à un modèle de description des connaissances. En effet, le savoir n'est pas un long couloir dans lequel on accède toujours à la même connaissance.

- Le labyrinthe maniériste d'Irrweg permet des choix alternatifs (Figure M.2 : Le labyrinthe d'Irrweg) : toutes les routes mènent à des points morts, sauf un qui est la sortie. Ce labyrinthe est un arbre de décisions, dans lequel les branches sont la représentation des décisions possibles.
- Enfin, le labyrinthe qui donne naissance aux réseaux et aux graphes est le « labyrinthe réseau » de Umberto ECO (Figure M.3 : Le labyrinthe réseau). Chaque point du labyrinthe peut être connecté à n'importe quel autre point. Cette structure a l'avantage d'être extensible à l'infini, il permet des connexions infinies et des corrections locales qui ne modifient pas le reste du labyrinthe. Évolutif, ce labyrinthe nécessite de l'utilisateur qu'il modifie en permanence l'image qu'il s'en fait : « Un réseau est un arbre auquel il faut ajouter des couloirs infinis connectant ses noeuds »⁸.

Modèle dans lequel les connexions, les liens, sont essentiels, le labyrinthe réseau permet une représentation multidimensionnelle des connaissances et un accès un à point précis par de multiples liens. En 1968, Ross QUILLIAN⁹ fait apparaître le réseau sémantique structuré conçu comme un réseau de noeuds interconnectés¹⁰. Le modèle qu'il décrit part d'un terme souche qui est défini par une série de noeuds, des tokens : ce n'est pour l'instant qu'un arbre. Seulement, les tokens peuvent à leur tour devenir des souches et porter des relations d'association : le réseau est ainsi constamment remodelé et modifié¹¹. Ce modèle en réseau permet alors la définition de chaque terme par ses connexions avec tous les autres termes ; il devient infini et multidimensionnel, non représentable en entier sur un plan bidimensionnel : la complexité du modèle-réseau peut seul être traité et compris par une machine.

L'apparition du modèle-réseau, du labyrinthe-réseau, a montré l'importance du lien dès la seconde moitié du XX^{ème} siècle. L'informatique a permis de mettre fin aux structures de connaissances qui étaient concevables par un esprit humain, afin de laisser la machine représenter les données dans toute leur complexité. Ce modèle, beaucoup plus efficace et riche que les arbres, consacre la valeur du lien, qui devient lui-même plus important que la donnée : il définit et légitime la donnée.

8. U. Eco, *De l'arbre au labyrinthe...*

9. Marvin Lee Minsky, *Semantic information processing*, Cambridge, Mass., 1968, p.227-270.

10. « *The memory model consists basically of a mass of nodes interconnected by different kinds of associative links* » in *Ibid.*, p.234

11. « *Token nodes make it possible for a word's meaning to be built up from other word meanings as ingredients and at the same time to modify and recombine these ingredients into a new configuration* » in *Ibid.*

7.2 Des labyrinthes de relations et d'identifiants : les hubs de liens

La modélisation des données sous la forme de labyrinthes — ou de graphes — a un impact considérable dans le Web de données : certains jeux de données sont eux-mêmes stockés dans une base de données graphe — comme Wikidata qui fonctionne sur la base de données Blazegraph— ; ou bien les jeux de données peuvent entre eux former un gigantesque graphe, infini. Cette seconde conception du modèle-réseau de Umberto ECO est au centre du Web de données. Avec la décentralisation des référentiels sur le Web et leur éclatement en de multiples données, il est apparu comme nécessaire de les recentraliser au travers de nouveaux référentiels fournisseurs d'un unique identifiant.

Cependant, la recentralisation passe également par l'ajout de données parallèlement à l'ajout des liens. En effet, nous l'avons montré au Chapitre 6 : Relier ses données à Wikidata : l'exemple de l'alignement des personnes physiques de l'INA, les données peuvent varier dans leur graphie et leur forme selon le référentiel duquel elles sont issues. Afin d'offrir des données utilisables par tous, certaines plateformes ajoutent, pour chacun de leurs identifiants, des données préférentielles aux côtés des liens pointant vers d'autres référentiels ou jeux de données.

7.2.1 De la décentralisation des référentiels à leur recentralisation dans le Web de données

La multiplication du nombre de référentiels dans le Web de données conduit à une profusion de données et à leur répétition, sans que soient repérées les données se rapportant à un même concept¹². L'importance prise par ces référentiels partagés, mis en commun, et réutilisés par tous a provoqué une recentralisation autour de nouveaux référentiels, nés de l'agrégation d'autres de ces référentiels.

Cette réaction, étonnante au premier abord puisqu'elle reconstruit des jeux de données alors que le mouvement inverse a été initié avec le Web de données pour plus d'efficacité, s'explique par la nécessité de créer davantage de liens entre les données et les référentiels¹³. Ces nouveaux référentiels constituent alors des « hubs » centralisant et ex-

12. La constellation du Linked Open Data montre cette augmentation croissante du nombre de référentiels dans le Web de données, et l'absence, pour certains, de liens vers d'autres référentiels. Voir Annexe I : La constellation du Linked Open Data (Figure I.1 : La constellation du Linked Open Data en juillet 2020).

13. « Ironie de l'histoire, alors que le Linked Open Data souhaitait mettre en relation des données hétérogènes et décentralisées chez différents fournisseurs, il aura suffi de 5 ans pour que les utilisateurs commencent à recentraliser leurs données au sein d'un espace unique » in G. Poupeau, *Au-delà des limites, que reste-t-il concrètement du Web sémantique ?...*

posant les données et les liens selon les règles et les principes du Web de données. Wikidata a ici un rôle majeur et prouve que une base de connaissances unique et centralisée est essentielle.

Si Wikidata est une base ouverte, des enjeux commerciaux peuvent également s'emparer de cette problématique de la structure des données sur le Web et de la description des documents nativement numériques : Google créé ainsi une base de connaissances parallèle — le Knowledge Graph —, elle aussi sous la forme d'un graphe mais n'utilisant pas RDF, utilisant Wikidata comme une source de données. De même qu'un langage universel était nécessaire pour décrire les documents sur le Web, un langage est nécessaire pour décrire les sites Web et leur offrir en échange un meilleur référencement : la puissance de Google, associé à Yahoo et Microsoft, lui permet ainsi d'imposer l'ontologie schema.org comme standard du Web pour la description des sites¹⁴, rejetant RDFa au profit de JSON-LD¹⁵.

Le Knowledge Graph et l'ontologie schema.org sont le meilleur exemple de l'utilisation idéale du Web de données et du Web sémantique. Ils ont été créés spécialement pour centraliser les données et y accéder par le seul langage machine. Wikidata, qui apparaît comme le référentiel le plus élevé dans le milieu bibliothéconomique¹⁶, n'est plus qu'une source de données structurées pour les robots des moteurs de recherche utilisant schema.org. Le référentiel a désormais acquis la place principale, centrale, dans l'administration de données et leur recherche.

7.2.2 Apparition des hubs de liens et d'identifiants

De même que les ontologies apportent une couche supplémentaire aux données en les liant entre elles malgré leurs différences, des hubs se forment pour centraliser les identifiants de différents référentiels afin d'avoir un unique lieu dans lequel puiser d'autres identifiants : c'est le principe de l'interopérabilité par suivi de liens. Ces nouveaux hubs se trouvant dans le Web de données, il est normal et indispensable qu'ils se créent sur les principes et les formats qui le composent. Ainsi, ces hubs forment des concepts, ou des entités, qui rassemblent les liens de vedettes et de concepts identiques. Ces nouveaux concepts sont également identifiés par des URIs servant d'identifiant pour chacun d'eux.

Ces hubs prennent une place centrale dans le Web de données et deviennent indispensables. Seulement, il peut être risqué de ne conserver dans un système documentaire qu'un seul identifiant, celui de ce hub. En effet, aucun de ces identifiants ou de ces liens — y compris les identifiants ark : de la Bibliothèque nationale de France (BnF) — ne peuvent

14. *Ibid.*

15. Une sérialisation en JSON adaptée aux données liées. Voir <https://www.w3.org/TR/json-ld/>.

16. Voir subsection 7.2.3 : Les hubs de liens et d'identifiants réceptacles de données.

être considérés comme pérennes et sûrs. Au-delà des identifiants qui disparaissent sur le Web, ce sont les données qui leur sont associées qui disparaissent¹⁷. Certes, l'ampleur de Wikidata peut laisser penser que les entités et leurs identifiants seront disponibles à long terme, mais des aléas techniques, financiers ou politiques peuvent très bien conduire à la fermeture de ce service et à la disparition de l'ensemble des données. Nous verrons par la suite (section 7.3 : Wikidata comme hub de liens : aligner les fictions et les séries de l'INA avec Wikidata) que la récupération de quelques identifiants majeurs est nécessaire en plus de l'identifiant de Wikidata, afin d'avoir à la fois un accès direct à ces identifiants — sans suivre les liens — et une relative sécurité quant à la conservation d'un lien avec un référentiel externe.

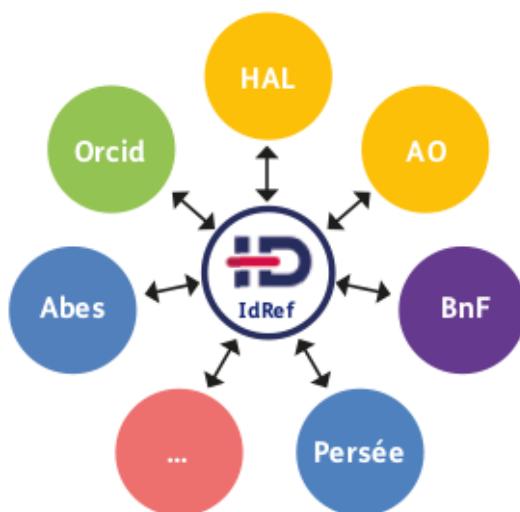


FIGURE 7.1 – La fédération entreprise par IdRef [Source : BOULET (Vincent), MISTRAL (François), ROUSSEAUX (Olivier), NICOLAS (Yann) et LE PAPE (Philippe), *Arabesques* n°85, réd. par David Aymonin, t. 85, Montpellier, 2017 (Arabesques), URL : <http://www.abes.fr/Publications-Evenements/Arabesques/Arabesques-n-85> (visité le 14/07/2020), p.9]

VIAF, IdRef ou LCSH sont quelques uns de ces hubs de liens qui fournissent en une seule page plusieurs identifiants. IdRef est basé sur les autorités du Sudoc, mais a été créé pour effectuer l'interopérabilité entre les différents catalogues du milieu de la recherche à travers un référentiel commun¹⁸. Ainsi, de multiples référentiels sont liés avec IdRef par leur identifiant : RAMEAU est utilisé, de même que HAL pour les publications scientifiques ouvertes ou Persée ; les identifiants VIAF et *International Standard Name Identifier* (ISNI) sont également renseignés, ... (Figure 7.1 : La fédération entreprise par

17. « De très nombreuses initiatives d'exposition des données ont aujourd'hui disparu, emmenant avec elles non seulement les identifiants mais aussi les données elles-mêmes » in *Ibid.*

18. « Il s'agissait de montrer la fonction de pivot des identifiants et les bénéfices d'un adossement des catalogues à un référentiel commun. » in V. Boulet, F. Mistral, O. Rousseaux, *et al.*, *Arabesques...*, p.9

IdRef)

7.2.3 Les hubs de liens et d'identifiants réceptacles de données

« D'un hub de références, Wikidata tend à devenir un réceptacle des données elles-mêmes¹⁹. »

Le projet Wikidata est né en 2012 de la volonté de centraliser les informations des Wikipédias sous la forme de données structurées en déclarations. Si le modèle de données ressemble à RDF, il ne l'est pas en raison des informations supplémentaires apportées par le modèle Wikidata : l'ajout de références et de qualificatifs enrichi les données. Wikidata s'impose dans le Web de données par ses atouts : ses données, enrichies par une communauté imposante, sont disponibles immédiatement et ne dépendent pas des mises à jour — comme cela était le cas avec DBpedia — ; les données de Wikidata se requêtent avec SPARQL.

Premièrement, Wikidata est un hub d'identifiants, de liens. Ces identifiants font l'objet, sur les pages d'entités de Wikidata, d'une partie à part, à la fin de cette page. S'ils sont mis à part, ils sont néanmoins de simples déclarations propriété-valeur avec une possibilité d'ajouter des qualificatifs et des références²⁰. L'ensemble — plusieurs centaines — des propriétés d'identifiants est répertorié sur la page de liste de ces propriétés Wikidata :List of properties/Wikidata property for an identifier : les jeux de données et les référentiels étant de plus en plus nombreux sur le Web, ces propriétés permettent de typer et de décrire chaque relation avec un identifiant.

Deuxièmement, Wikidata est un réceptacle de données. En effet, en plus des identifiants, Wikidata propose des déclarations structurées biographiques ou générales sur l'entité. En cela, si IdRef et Wikidata semblaient être similaires avec l'offre de liens, ils diffèrent par cette structuration permanente des informations. Ainsi, le concept Vincent DEDIENNE de IdRef²¹ structure l'état civil et les dates, mais met le reste des informations en note : « Comédien,auteur , acteur et humoriste français ». Wikidata apporte du sens grâce à la propriété P106 qui permet d'indiquer que la déclaration concerne l'*occupation* de la personne, et d'offrir un lien vers l'entité de cette occupation.

Par la structuration de ses données et l'apport de multiples identifiants et liens vers d'autres jeux de données et référentiels, Wikidata devient un *super hub*, se plaçant plus

19. G. Poupeau, *Au-delà des limites, que reste-t-il concrètement du Web sémantique ?...*

20. Un compte Twitter, qui a un identifiant et qui peut par conséquent faire l'objet d'une déclaration dans une entité de Wikidata, a comme qualificatif le nombre d'abonnés à une date donnée.

21. Vincent DEDIENNE dans IdRef : <https://www.idref.fr/196914183>

Vincent DEDIENNE dans Wikidata : <https://www.wikidata.org/wiki/Q18413745>

haut encore que ceux évoqués précédemment.

L'agrégation de liens et d'identifiants est un enjeu essentiel de manière à n'effectuer des opérations d'alignement qu'une seule fois, et ainsi de diminuer le coût de ces opérations. Ces agrégations nécessitant elles-mêmes des identifiants pour pouvoir exister sur le Web de données, des données structurées sont venues s'ajouter à la simple exposition de liens vers d'autres référentiels. Ainsi, Wikidata a montré son efficacité et se place aujourd'hui comme acteur principal d'agrégation de connaissances et de liens vers d'autres sources de données souvent plus spécialisées.

7.3 Wikidata comme hub de liens : aligner les fictions et les séries de l'INA avec Wikidata

Le principal intérêt de Wikidata, nous l'avons évoqué précédemment, est l'agrégation de liens et d'identifiants d'autres référentiels et jeux de données. Cela permet d'accéder à un même endroit à divers identifiants de référentiels, sans avoir à effectuer un alignement avec chacun de ces référentiels. L'opération d'alignement de fictions et de séries avec Wikidata ne peut s'effectuer qu'à partir des titres : alors qu'un alignement de personnes se réalise sur un voire deux mots de l'état civil, celui de titres et de séries doit se réaliser sur l'ensemble des mots de ces titres, ce qui augmente considérablement les échecs d'alignement.

Cependant, malgré les difficultés, imposées une nouvelle fois par le langage naturel, l'alignement reste une opération importante pour l'enrichissement des données d'une institution : le parcours de liens devient alors possible, et l'accès au Web de données apporte de nouvelles informations sur les instances alignées.

7.3.1 Enrichir ses données avec des identifiants plutôt qu'avec des textes

L'établissement de liens avec des référentiels externes est essentiel à l'INA. En effet, les données issues du DL sont parfois sommaires et tournées vers l'événement de diffusion au détriment de la description documentaire. Si les données achetées à l'extérieur — auprès de Plurimédia, Médiamétrie, ... — apportent les informations les plus importantes, l'ensemble des acteurs d'une série ou d'une fiction ne sont par exemple pas présents dans les bases de données de l'INA.

Une première possibilité serait alors d'aligner les fictions et les séries avec Wikidata afin de récupérer les libellés des valeurs de la propriété P161 qui permet la déclaration

de membres du casting. Cette possibilité nécessiterait néanmoins une mise à jour régulière avec un nouveau lancement de l'alignement, de manière à avoir les données et les informations actualisées de Wikidata. De plus, l'introduction de ces textes coupe les liens qui étaient présents sur Wikidata, et empêche ainsi de naviguer de lien en lien depuis un membre de casting, par exemple, pour arriver sur une autre de ses fictions.

Le stockage du seul identifiant de l'entité Wikidata de la fiction ou de la série suffit alors. À partir de cet identifiant, il devient possible d'accéder à toutes les déclarations de l'entité, qu'elles soient biographiques ou générales, ou bien qu'elles soient des liens, ainsi qu'aux déclarations des valeurs de ces déclarations, ... L'alignement des fictions et des séries de l'INA vise donc conséquent à récupérer les identifiants Wikidata des fictions, des séries, et des épisodes de séries.

Parallèlement à cette récupération d'identifiants Wikidata, il est également nécessaire d'obtenir l'ISAN, l'identifiant international de tout document audiovisuel. Cet ISAN permet alors d'obtenir par rebond des informations plus précises sur la fiction ou la série grâce à une base de données spécifique, IMDb²². Le stockage à l'INA de simples identifiants — Plurimédia, Médiamétrie, IMedia, Wikidata et ISAN — est alors suffisant pour avoir accès au Web de données et à l'ensemble des informations et des données disponibles sur ces instances.

7.3.2 Aligner des fictions et des séries avec Wikidata

De même que l'alignement des personnes physiques avec Wikidata²³, le SPARQL-EndPoint et l'API Wikibase sont conjointement utilisés. Une première étape nécessite la récupération de l'ensemble des sous-classes des entités *Film* (Q11424)²⁴ et *Série Télévisée* (Q5398426)²⁵ par une requête SPARQL (Figure 7.2 : Requête SPARQL pour récupérer les sous-classes de l'entité *Série Télévisée*). Dans le cas des séries télévisées, d'autres en-

```
select ?serie ?serieLabel
where{
?serie wdt:P279* wd:Q5398426.
service wikibase:label{bd:serviceParam wikibase:language "fr,en"}
}
```

FIGURE 7.2 – Requête SPARQL pour récupérer les sous-classes de l'entité *Série Télévisée*

tités, non présentes dans la requête de la Figure 7.2 : Requête SPARQL pour récupérer

22. **noauthor_imdb_nodate**.

23. Voir Chapitre 6 : Relier ses données à Wikidata : l'exemple de l'alignement des personnes physiques de l'INA.

24. Film (Q11424) : <https://www.wikidata.org/wiki/Q11424>

25. Série Télévisée (Q5398426) : <https://www.wikidata.org/wiki/Q5398426>

les sous-classes de l’entité *Série Télévisée*, sont nécessaires pour avoir accès à l’ensemble des séries télévisées de Wikidata. Il est ainsi nécessaire de faire une seconde requête avec la propriété P361 qui indique qu’une entité est « une partie d’ » une autre entité : l’entité de la *saison* (Q3464665)²⁶, essentiel dans l’alignement, est récupéré par ce moyen.

La récupération des instances de chacune des sous-classes est ensuite possible par une requête SPARQL²⁷ (Figure 7.3 : Requête SPARQL pour récupérer les identifiants des instances de la classe *Saison*). L’objectif de cette étape n’est pas de récupérer l’ensemble des valeurs des propriétés qui permettent l’alignement — ce qui ne serait pas possible en raison des limites du SPARQL-EndPoint évoquées au Chapitre 6 : Relier ses données à Wikidata : l’exemple de l’alignement des personnes physiques de l’INA —, mais d’obtenir les identifiants de toutes les entités instances des sous-classes.

```
select ?saison ?saisonLabel
where{
?saison wdt:P31 wd:Q3464665.
service wikibase:label{bd:serviceParam wikibase:language "fr,en"}
}
```

FIGURE 7.3 – Requête SPARQL pour récupérer les identifiants des instances de la classe *Saison*

L’API Wikibase, avec le module *wbgetentities* permet ensuite d’obtenir les points de comparaison avec les données de l’INA :

- pour les fictions, les valeurs suivantes sont ainsi récupérées :
 - P577 pour la date de publication de la fiction
 - P57 pour le réalisateur
 - P162 pour le producteur
 - les libellés préférentiels et alternatifs du titre et des noms de personnes sont également ajoutés
- pour les séries, plus précisément les épisodes, les suivantes :
 - P577 pour la date de publication de l’épisode
 - P179 pour la série d’appartenance de l’épisode
 - le qualificatif P1545 de P179 pour le numéro de l’épisode
 - P4908 pour le nom de la saison
 - le qualificatif P1545 de P4908 pour le numéro de la saison
 - les libellés préférentiels et alternatifs des titres sont récupérés en français et en anglais — en effet, un grand nombre de séries conservées à l’INA n’ont pas de titres traduits

26. Saison (Q3464665) : <http://www.wikidata.org/entity/Q3464665>

27. Le SPARQL-EndPoint est ici utilisé en raison du faible nombre de requêtes qui seront effectuées, et du nombre de résultats par requête relativement faible (inférieur à 200000) n’induisant pas de *timeout*.

- les libellés préférentiels et alternatifs des valeurs des propriétés sont récupérés uniquement en français

Cette récupération de différentes données permet un alignement avec les données de l'INA. La fiction long métrage « Doux dur et dingue » de James FARGO en 1978 trouve ainsi son entité équivalente (Q1195524) dans Wikidata grâce à une stricte égalité — après passage des majuscules en minuscules et suppression de la ponctuation — des titres : l'ISAN (déclaré avec la propriété P3212) « 0000-0000-3B9A-0000-D-0000-0000-Z » peut alors être ajouté aux données de l'INA en plus de l'identifiant Wikidata. Pour les séries, ce sont les épisodes qui subissent l'alignement car chacun d'entre eux est identifié individuellement et lié ensuite avec sa série d'appartenance : le troisième épisode de la comédie de situation « 3ème planète après le Soleil », « The Fifth Solomon », est alors aligné doublement. Il l'est une première fois avec sa série (Q870490), et une seconde fois avec son entité équivalente dans Wikidata, Q18040623. Enfin, l'ISAN, quand il est disponible, est également ajouté, ce qui est le cas pour cet épisode identifié internationalement par l'identifiant « 0000-0001-637C-0065-4-0000-0000-P ».

7.3.3 Les difficultés posées par les langages naturels

Les résultats obtenus après un alignement de fictions ou de séries avec Wikidata peuvent paraître faibles : si les fictions sont alignées presque à 50%, ce n'est pas le cas des séries pour lesquelles à peine 10% des épisodes ont pu être alignés avec leur équivalent Wikidata. Bien que de nombreux épisodes et séries n'existent pas sur Wikidata, comme « Mon père dort au grenier » dont il existe 26 saisons, cette absence d'entités Wikidata n'est pas suffisante pour expliquer les faibles alignements des séries.

La raison principale est le langage naturel utilisé de chaque côté de l'alignement, et les nombreuses divergences de graphie qui peuvent exister sur les mots des titres. En effet, les titres des séries sont pour certains traduits en français, d'autres restent en anglais, à la fois dans les données de l'INA et sur Wikidata. L'alignement n'utilisant pas de traducteur ou d'intelligence artificielle, la similarité entre, par exemple, l'instance de l'INA « Monk va à la noce » de la saison 7 de Monk, avec l'entité Q50846176 « Mr. Monk Goes to a Wedding » qui lui correspond. L'alignement n'aura alors réussi à aligner que la série d'appartenance (Q189068) de cet épisode, et non l'épisode lui-même. L'utilisation des libellés préférentiels et alternatifs en français et en anglais aura, ici, été inutile. Cependant, la prise en compte de l'anglais a permis de nombreux alignements d'épisodes qui, tant du côté de l'INA que de Wikidata, n'ont pas été traduits.

De plus, des séries très longues, comme « Amour, gloire et beauté », qui compte plus de 5000 épisodes, peuvent ne pas être décrites dans Wikidata au niveau de l'épisode.

Quand une entité d'épisode est disponible dans Wikidata et que ce titre ne correspond pas à celui de l'INA, il est possible d'utiliser le numéro de saison et d'épisode pour effectuer l'alignement. Seulement, le comptage des épisodes est différent selon Wikidata et l'INA : Wikidata ne réinitialise pas le numéro d'épisode au début de chaque saison, alors que l'INA le fait le plus souvent.

Les difficultés à l'alignement des fictions et des séries sont multiples, ce qui entraîne un faible rendement, notamment quand il s'agit d'aligner à la fois un libellé et un autre libellé d'une valeur de déclaration. Les graphies et les langues varient énormément. Les limites d'un alignement par stricte égalité sur des textes sont certainement ici atteintes. Des méthodes alternatives seraient nécessaires comme l'utilisation d'un traducteur puis de réseaux de neurones permettant d'aligner sur des similarités de textes et des égalités de numéro de saison ou de producteur.

Bien que nécessaire, la récupération d'identifiants et de liens sur le Web de données, principalement sur Wikidata, peut se révéler difficile selon le type de données permettant l'alignement, et le genre des entités et des instances. L'alignement de personnes, effectué sur peu de mots et des données structurées comme les dates, est ainsi plus facile à réaliser que l'alignement de séries qui ne peut se faire qu'à partir d'une longue chaîne de caractères, le titre.

Le succès des modèles de données en graphe est incontestable et est repris dans tous les projets d'envergure internationale : nous avons évoqué Wikidata, le projet European Holocaust Research Infrastructure (EHRI) ²⁸ peut également être cité comme acteur institutionnel se dégageant des bases de données relationnelles et des traditionnels référentiels hiérarchiques et contrôlés. Cependant, ce modèle de graphe nécessite un langage de requête efficace et des rendements élevés. Or, il a été constaté avec Wikidata et le SPARQL-EndPoint des lenteurs de retours de résultats, obligeant à utiliser d'autres moyens pour obtenir les entités de Wikidata avec leurs déclarations.

La recentralisation des données autour de quelques acteurs du Web de données est une conséquence inattendue de la décentralisation de ces mêmes données qui avait eu lieu quelques années auparavant avec la publication de jeux de données et de référentiels sur le Web sémantique. Cette recentralisation est née d'un besoin d'obtenir en un même endroit une multiplicité de liens, d'identifiants et d'informations, sans avoir à connaître

28. Site du projet : *European Holocaust Research Infrastructure*, URL : <https://ehri-project.eu/> (visité le 16/09/2020)

Utilisation du graphe de données : Tobias Blanke, Michael Bryant et Reto Speck, “Developing the collection graph”, *Library Hi Tech*, 33–4 (2015), p. 610-623

les institutions ou les jeux de données dans lesquels aller chercher les données nécessaires. Si cette recentralisation concerne ici le Web de données, elle peut également concerner les institutions directement, ainsi que les entreprises : on ne parle alors plus de Linked Open Data, mais de Linked Enterprise Data (LED).

Chapitre 8

Le *Lac de données* de l'INA : le référentiel au centre du modèle

L'impact du LOD sur la structure des données et l'éclatement des référentiels a été l'un des aspects de la refonte des systèmes d'information en institutions ou en entreprises, sous la forme de LED. Ces LED sont des modèles de données à la structure similaire au LOD, ce qui les rend très efficaces et utiles dans l'utilisation des données qui en découlent.

Le *Lac de données* de l'INA est l'un de ces LED : il a vocation à regrouper l'ensemble des métadonnées de l'Institut, en provenance de plusieurs départements sous diverses formes. L'opération de traitement qui est nécessaire pour la création de ce *Lac* permet de les enrichir par de multiples liens avec le LOD ou des référentiels internes : le lien devient une notion prioritaire et essentielle entre des données d'un même document qui sont éclatées en plusieurs instances ou concepts.

8.1 Application des principes du Web de données aux systèmes documentaires : le Linked Enterprise Data

L'apport du Web de données à la structure des données et à la place des référentiels est important. Lier un système documentaire au Web de données est possible ; repenser ce système documentaire selon les principes du Web de données pour s'adapter aux nouveaux besoins et aux nouveaux usages est une pratique de plus en plus courante dans les institutions patrimoniales. L'INA a ainsi entrepris une réflexion sur cette transformation dès 2014, et sa mise en œuvre en 2015. Face à l'accumulation de bases de données, un modèle de données inspiré du Web de données doit pouvoir recentraliser les métadonnées et assurer l'interopérabilité de l'ensemble du système documentaire.

L'interopérabilité du système — uniquement celui de l'INA, il n'est pas nécessaire ni envisageable de le rendre interopérable avec les autres institutions et le Web de données — est seulement possible par le repositionnement du référentiel en son sein.

8.1.1 Permettre l'interopérabilité au sein des institutions

L'interopérabilité du système documentaire est le principal enjeu du *Lac de données*. Nous l'avons évoqué au Chapitre 3 : Les référentiels à l'INA, l'INA possède plusieurs bases de données distinctes, propres à chaque métier et à chaque besoin. Les référentiels ne sont communs qu'entre les métiers aux mêmes besoins : la DJ et la DDCOL ne partagent pas le même référentiel de personnes physiques et morales ; il existe par conséquent deux de ces référentiels au sein d'une même institution.

La création d'un LED permet de centraliser ces référentiels et de les partager entre les différents corps de métier, qu'ils soient juridiques, patrimoniaux ou commerciaux. Elle permet également de casser le monolithisme¹ du système documentaire, pensé comme un tout répondant à un besoin à un instant précis. Seulement, l'évolution des usages, l'évolution des usages, et l'évolution des documents à décrire, entraînent une modification de la description qui est réalisée, et par conséquent une sédimentation de rajouts aux bases de données sources. En effet, étant conçues pour un unique besoin, ces bases supportent mal les modifications de modèle de données, ce qui crée de nouveaux attributs dans les tables, ou bien de nouvelles tables dans les bases de données.

Les difficultés posées par ces multiples bases de données concernent également l'utilisation qui est faite des métadonnées. De même que des portails sont des moyens d'assurer une interopérabilité — par plus petit dénominateur commun — entre deux jeux de données du Web de données, l'application Hyperbase de l'INA permet de consulter les données des bases DA et DL². Mais cette application comble seulement partiellement des différences de modélisation des données dans chacune des bases de données. La création de cette application répondait au besoin de pouvoir consulter sur une même page des données provenant de diverses bases. De nombreuses autres applications ont été créées pour répondre rapidement, chacune, à un besoin : Totem pour le DA ou MediaIndex pour le DL sont deux exemples de ces applications aux usages similaires propres à chaque métier.

1. Terme employé par Emmanuelle BERMÈS, Gautier POUPEAU et Antoine ISAAC dans E. Bermès, A. Isaac et G. Poupeau, “Cas D : Lier les données internes avec le LED”, dans *Bibliotheques*, ISSN : 0184-0886, 2013, p. 153-164, URL : <https://www-cairn-info.proxy.chartes.psl.eu/le-web-semantique-en-bibliotheque--9782765414179-page-153.htm> (visité le 01/08/2020)

2. Voir Annexe D : Les bases de données de la DDCOL de l'INA (Figure D.1 : Les bases de données de la DDCOL de l'INA).

L'utilisation d'un LED permet de retourner l'utilisation qui est faite d'un système documentaire : plutôt que de partir des besoins et des usages qui seront faits des données, la réflexion se porte d'abord sur les données afin de bâtir un modèle de données qui puisse s'adapter à l'évolution des besoins, sans avoir besoin de les prévoir. Le LED rend leur cohérence aux données et aux informations, permet une meilleure gestion de ces données et informations, et une amélioration des services rendus à l'utilisateur final.

8.1.2 Repenser le système documentaire

L'objectif du LED est l'interopérabilité, la connexion entre les jeux de données de l'institution, ou de l'entreprise, qui ont des structures différentes mais partagent des points communs comme les référentiels. Pour cela, les processus ETL (Extract-Transform-Load) sont essentiels. De multiples bases de données, l'objectif est d'en obtenir une seule en conservant la totalité des données migrées. Au cours de ce traitement pour restructurer chaque donnée, il est possible d'apporter un enrichissement au travers d'alignements avec d'autres jeux de données ou référentiels. Ces alignements, nous l'avons montré, peuvent être de deux types :

- internes, entre deux jeux de données de l'institution³ pour assurer l'interopérabilité du système
- externes, entre un jeu de données de l'institution et un jeu de données d'une autre institution grâce au Web de données⁴

En repensant le système documentaire depuis les données au lieu des besoins et des usages qui en seront faits, de multiples usages peuvent naître et sont facilités dans leur développement par la centralisation des données : à l'INA, le *Lac de données* permet d'alimenter plusieurs applications et sites Web, comme ina.fr, madeleina.ina.fr ou inamediapro.fr. De même que dans le Web de données, chaque document, chaque instance de l'INA et du LED se voit attribuer un identifiant unique, facilitant ainsi l'établissement de liens entre les instances, ou avec les référentiels. Ces identifiants permettent une interopérabilité par les liens, similaire au Web de données : ainsi, l'interopérabilité du LED ne passe pas, comme cela pouvait être le cas en bibliothèque avec le format MARC, par une interopérabilité par un format unique.

8.1.3 Le positionnement du référentiel

La création des liens entre les instances du LED nécessite, lors du processus d'ETL, d'utiliser des référentiels ou de trouver les points de contacts entre les jeux de données.

3. Voir section 2.3 : Passer du texte libre à un vocabulaire contrôlé : aligner des notes qualité et un thésaurus de noms communs.

4. Voir Chapitre 6 : Relier ses données à Wikidata : l'exemple de l'alignement des personnes physiques de l'INA et section 7.3 : Wikidata comme hub de liens : aligner les fictions et les séries de l'INA avec Wikidata.

Ces référentiels deviennent des pivots dans le système documentaire : le DA et le DL partagent des structures de données différentes ; pourtant, des liens entre ces deux jeux de données peuvent être établis grâce aux référentiels — celui des personnes physiques et morales, celui des types de matériels, le thésaurus des noms communs, ... Utiliser des référentiels pour établir du lien ne nécessite pas d'alignement entre les données puisque les tables des bases de données sont déjà liées aux référentiels. En revanche, l'utilisation des points de contact nécessite la réalisation d'alignements, de manière à recoller deux mêmes concepts ou instances de deux jeux de données ou de deux référentiels.

La création d'un référentiel commun dans le LED apparaît par conséquent nécessaire et indispensable. Cependant, ce référentiel peut ne pas être créé spécifiquement pour le LED, mais être une réutilisation d'un référentiel existant dont on aurait décidé de manière commune que sa valeur est supérieure à un autre : dans le cas des référentiels des personnes physiques de l'INA, un choix doit être fait pour décider lequel des référentiels de personnes de la DDCOL ou de la DJ imposera ses normes de graphie. De même, des choix doivent être effectués quant aux référentiels externes utilisés : Wikidata est une base de connaissances globale comprenant plusieurs dizaines de millions d'entités, mais cette base n'est pas assez complète pour pouvoir être utilisée à l'INA comme référentiel commun sur lequel l'ensemble des données repose. En effet, nous l'avons montré lors de l'alignement des personnes physiques avec Wikidata, de nombreuses personnes, peu ou pas connues, ne font pas l'objet d'une entité Wikidata : l'INA ne peut alors qu'enrichir ses données avec l'identifiant de Wikidata, ou avec d'autres identifiants extraits de Wikidata grâce au hub de liens et d'identifiants qui représente cette base de connaissances. Le repositionnement des référentiels au centre du système documentaire permet ainsi d'éviter les redondances de données entre les bases.

Cependant, nous le verrons ensuite (section 8.2 : Le *Lac de données* de l'INA : un repositionnement du référentiel au centre du modèle de données), la présence d'un référentiel défini comme référentiel n'est pas indispensable. En effet, comme dans le Web de données, le référentiel s'est progressivement disloqué en données avec les modèles en graphe, faisant alors de chaque référentiel un jeu de données comme les autres. Plus encore, un jeu de données qui n'est pas défini comme référentiel peut à son tour devenir référentiel s'il est utilisé et lié avec une autre donnée.

L'impact du Linked Enterprise Data est multiple, mais contribue notamment à posséder une base de données cohérente et structurée, laquelle ne dépend pas des utilisations qui en sont faites. Ainsi, une application utilise la base de données et sa structure, sans avoir d'incidence sur le stockage et la gestion des données. Le LED permet alors une grande évolutivité du système dans les modifications qui lui sont apportées.

8.2 Le *Lac de données* de l'INA : un repositionnement du référentiel au centre du modèle de données

La fusion du DL et du DA dans la DDCOL en 2012 n'ont pas permis de centraliser les deux silos de données existants. Ainsi, en 2015, le projet du *Lac de données* est lancé afin de centraliser les données et les métadonnées de tout l'Institut, afin de les mettre en cohérence, de supprimer les redondances et les barrières techniques ou structurelles de ces données, de répondre enfin aux nouveaux enjeux de la fin des années 2010.

Un nouveau modèle de données, construit sur le contenu intellectuel et non les usages, et accompagné d'une nouvelle infrastructure centralisée pour l'accueil des données et des métadonnées, voit le jour. Le référentiel, déconstruit, devient une donnée identique aux données résultant de la décomposition de la notice documentaire. Au-delà de ce modèle de données, c'est l'ensemble du système d'information qui subit cette refonte, avec un stockage des données, leur traitement et des accès repensés.

8.2.1 Le *Lac de données* , un modèle basé sur des classes d'entités

Le Web de données avait permis la déconstruction des informations en données, l'effacement du référentiel au profit de données déstructurées mais liées. La réflexion sur le nouveau modèle de données de l'INA a conduit au même phénomène : les données bibliographiques — de description des documents audiovisuels — se retrouvent au même niveau que les données d'autorités issues des référentiels. Les données d'autorités ne se trouvent plus à la marge du système documentaire, mais bien intégrées dedans, au centre, puisqu'elles deviennent indispensables dans la description des documents.

Le modèle de données du *Lac de données* propose une structure globale, et adaptable à chaque besoins non pensés lors de la modélisation, pour accueillir les données. Pour cela, le modèle du *Lac de données* est basé sur les relations entre cinq principales classes d'entités⁵. En cela, ce modèle peut ressembler aux FRBR⁶ avec ses quatre grandes classes item, manifestation, expression et œuvre ; le modèle du CIDOC-CRM ou d'autres modèles à entités peuvent également être comparés. Le *Lac de données* n'adopte aucun des modèles de données existant dans le domaine bibliothéconomique ou patrimonial en raison de la spécificité des fonds conservés et des données documentaires produites, ainsi

5. Voir Annexe J : Repenser la place du référentiel (Figure J.3 : Le modèle de données du *Lac de données* à l'INA).

6. Présentés précédemment dans la section 4.1 : Le web de données : naissance et principes.

que de la singularité de l'historique de la DDCOL qui conserve à la fois des données orientées événement ou archivistiques.

L'**instance** est la première des cinq entités principales du *Lac de données*. Elle correspond à l'entité intellectuelle du contenu — un programme, qu'il soit par exemple une émission, ou bien un reportage ; une photographie ; une documentation d'accompagnement, un épisode de série ou une fiction, ...) : sans l'être exactement, l'instance peut être rapprochée de l'œuvre des FRBR.

L'**événement** permet la description d'un événement attaché à une instance : cet événement peut être lié à la création du contenu (la captation, l'enregistrement, la prise de vue pour une photographie, ...), à l'exploitation qui en a été faite (diffusion télévisuelle, projection des *Actualités françaises* au cinéma, mise en ligne sur l'une des plateformes de l'INA, ...), ou bien à l'archivage et à l'usage de ce contenu (numérisation, restauration, description des droits et des informations juridiques pesant sur le contenu, ...).

L'**item** correspond au matériel — physique ou numérique — sur lequel se trouvent les contenus (bandes LTO du dépôt légal, photographie, ...).

Les **textes** permettent une description du contenu par le langage naturel : il peut s'agir de titres extraits de génériques ou saisis par le technicien de gestion des collections multimédia lors du catalogage ; les identifiants sont également des textes ; ... Ces textes ne sont pas soumis aux référentiels et ne les constituent pas.

Le **concept** est le pivot du modèle de données puisqu'il représente les référentiels. Il permet la description de toutes les instances.

L'identification de chacune de ces entités est nécessaire puisque le modèle de données du *Lac de données* est conçu à partir de relations entre ces entités : le Web de données dispose d'URIs HTTP, ce modèle de données d'une entreprise utilise des identifiants non significatifs (il s'agit d'une suite de douze chiffres et lettres) pour créer du lien au sein de son système documentaire (Figure 8.1 : Modélisation des cinq entités du *Lac de données*).

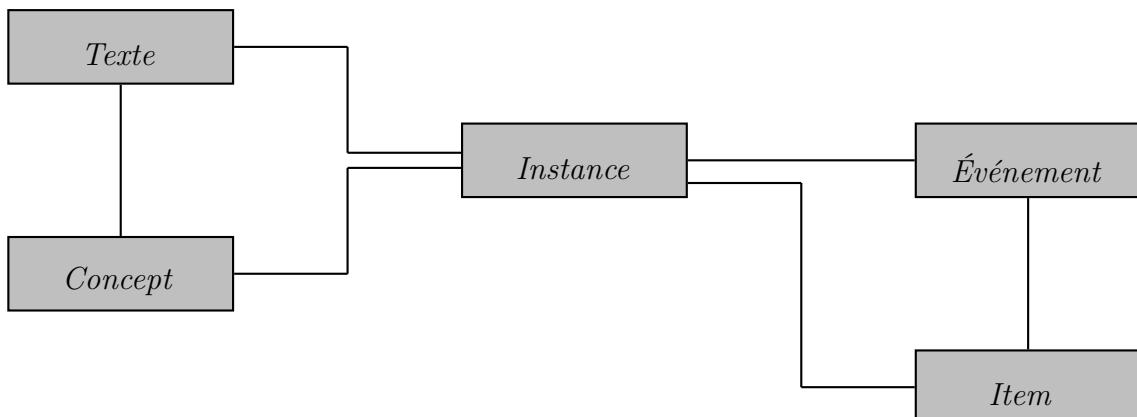


FIGURE 8.1 – Modélisation des cinq entités du *Lac de données*

8.2.2 La place des concepts

Le modèle de données établi dans le *Lac de données* permet de ne créer qu'un seul « référentiel » grâce aux concepts. Ces derniers permettent la description de l'ensemble des instances : dans ce modèle, un grand nombre de données sont comprises comme des concepts. C'est pourquoi ils regroupent une grande variété de noms communs et de noms propres dont :

- les personnes physiques et morales, tirées du référentiel des personnes physiques et morales de la DDCOL
- les noms communs issus du thésaurus des noms communs de cette même DD-COL : ils offrent les autorités matière nécessaires à la description (l'instance concerne-t-elle le sport ? la télévision ? la cuisine ? ...)
- le genre de l'instance (émission, reportage, film, épisode de série, fiction, ...)
- la provenance de l'instance, ce qui concerne notamment les codes des chaînes de télévision et des stations de radio employés dans les bases sources et repris dans le *Lac de données*, mais concerne également la base source de provenance des données du *Lac de données*. En effet, la migration de données depuis une base vers une source nécessite de conserver la trace de son parcours afin de repérer d'éventuelles erreurs de mapping : ainsi, les codes des bases sources sont ainsi des concepts.
- etc.⁷

Les entités du *Lac de données* sont similaires aux entités de Wikidata par la nécessité de la création de liens entre elles afin qu'elles puissent exister dans le modèle de données. Afin de relier ces cinq entités et de mettre en cohérence les données, des relations typées sont créées entre les entités, notamment entre les concepts, de manière à identifier le rôle, le type, ou la fonction du concept par rapport à l'instance ou au texte. Ainsi, des tables permettent, comme pour les annotations ou les crédits, de lier des concepts aux instances afin d'apporter du sens. De plus, il est possible avec ce modèle de données d'établir une relation entre deux concepts, afin d'exprimer par exemple la provenance d'un concept d'une personne (Figure 8.2 : Modélisation globale des relations entretenues par les concepts).

L'établissement de relations entre les entités n'est possible qu'avec les identifiants attribués à chacune des entités. Un concept n'est par conséquent pas défini dans un seul endroit, à une seule table : un graphe de relations se met en place tout autour de lui afin de le définir le plus précisément et de lui apporter du contexte et du sens. Ces relations internes sont essentielles au fonctionnement du système documentaire et de la cohérence(Figure 8.3 : Modélisation du concept Mylène FARMER dans le *Lac de données*).

7. Plusieurs millions de données sont devenus des concepts dans le nouveau système documentaire de l'INA.

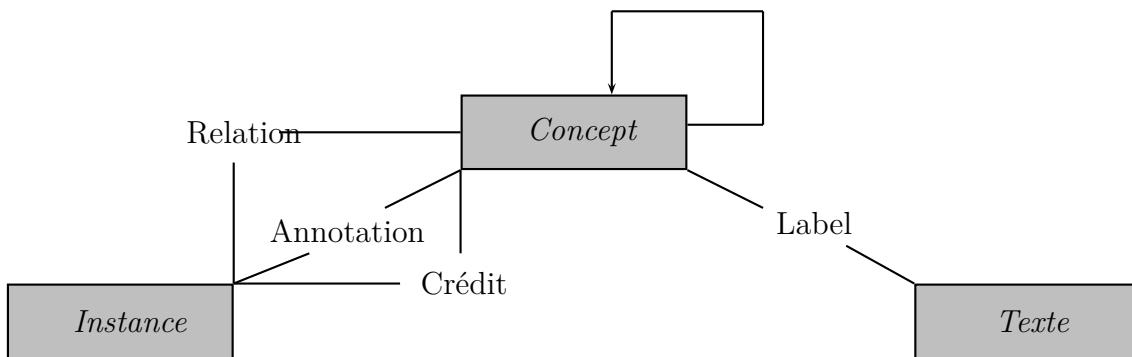


FIGURE 8.2 – Modélisation globale des relations entretenues par les concepts

Le modèle de données du *Lac de données* lui permet également d’obtenir une ouverture à l’extérieur, vers le Web de données, afin d’obtenir des informations et des données supplémentaires. Ainsi, la simple conservation de quelques identifiants comme ceux de Wikidata ou de la BnF, et des identifiants internationaux comme l’ISAN suffisent à créer des ponts avec le Web de données(Figure 8.4 : Modélisation du concept Mylène FARMER dans le *Lac de données* et le LOD).

8.2.3 Le *Lac de données* comme un LED : une infrastructure unique

« À l’heure où nous cherchons à faire fructifier la donnée, comme actif de l’entreprise, il est essentiel pour réussir justement à faire émerger de nouveaux usages de décloisonner nos silos de données et de libérer la donnée de l’usage pour lequel elle a initialement été créée.⁸ »

Le *Lac de données* n’est pas seulement la refonte d’un modèle de données. Afin de disposer des capacités de stockage, de traitement, et d’accès nécessaires, le *Lac de données* est également la création d’une infrastructure centralisée, depuis laquelle les applications futures pourront être créées : en cela, il est un LED, pensé depuis le bas, depuis les données, afin de permettre une multiplicité d’applications⁹.

La couche la plus basse de ce LED est la base de données. Le choix de celle-ci est essentiel afin de lier performance et modèle de données. Ainsi, chaque type de base de données ayant ses propres caractéristiques, ses propres avantages et ses limites, l’INA utilise les quatre types de bases de données en y dupliquant le modèle de données. Alors, le modèle de données est respecté et les applications nécessitant de la part des bases de

8. G. Poupeau, *Réflexions et questions autour du Web sémantique...*

9. Voir Annexe K : Repenser l’infrastructure (Figure K.1 : Schéma du *Lac de données* depuis le stockage jusqu’à l’accès aux données).

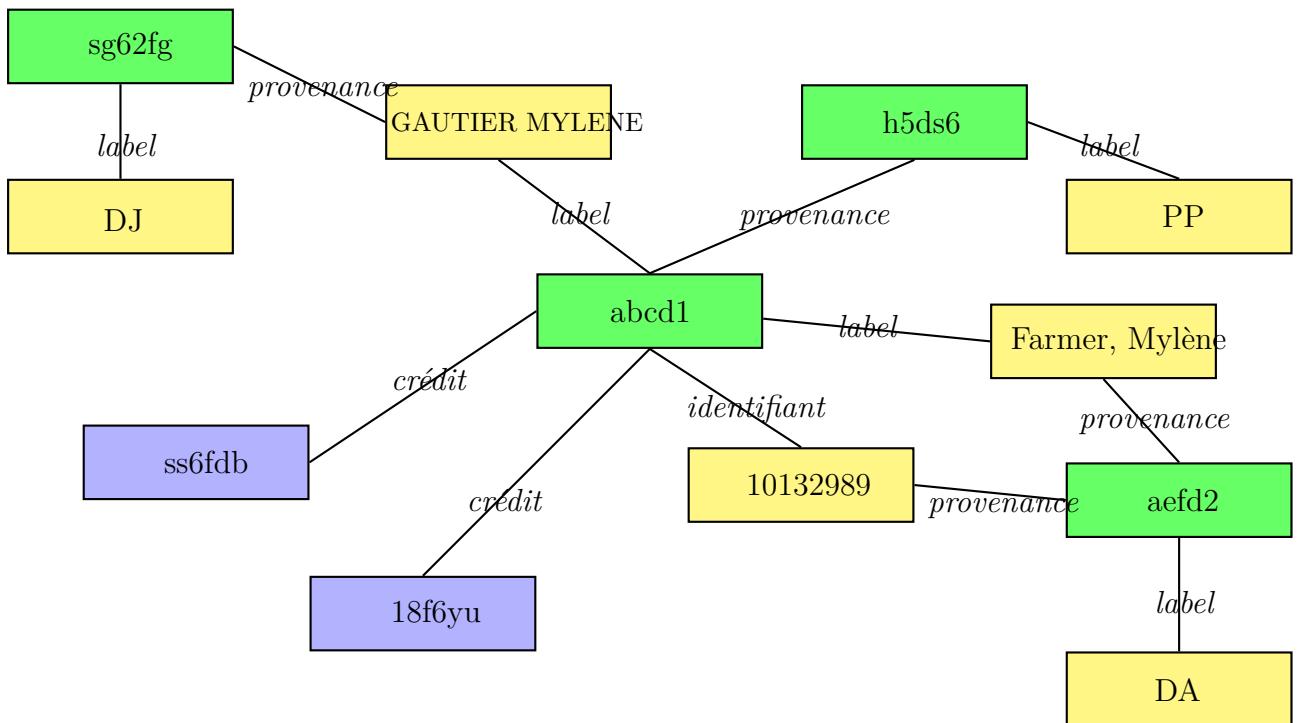


FIGURE 8.3 – Modélisation du concept Mylène FARMER dans le *Lac de données* [Données partielles d'exemple. Vert : concept. Jaune : texte. Bleu : instance.]

données de grandes performances pourront en utiliser une plutôt qu'une autre :

- les bases de données relationnelles offrent une forte structuration de la donnée, et de bonnes performances d'écriture et de lecture — ce qui avait conduit à leur adoption par le DA, le DL ou la DJ. Cependant, la création de relations entre les entités nécessite la création de nombreuses tables et l'éparpillement de la donnée dans la base ; les calculs nécessaires pour rassembler les données d'une entité deviennent complexes et longs, ce qui rend la base de données relationnelle peu performante pour le *Lac de données*
- la base de données document permet quant à elle une montée en charge rapide et importante avec la possibilité de conserver de grandes masses de documents, mais ne permet pas le respect de la structuration des données
- le moteur de recherche permet également cette montée en charge, ainsi qu'une recherche plein texte efficace et rapide obtenue par l'indexation des données dans le moteur
- enfin, la base de données graphe permet une structuration très fine des données obtenue par les liens établis entre ces données ; cependant, de même que les bases de données relationnelles, la performance lors des requêtes est très vite limitée dès que les requêtes se complexifient et ont pour but de restituer l'intégralité des informations concernant un document, un concept, ...

Pour tirer les avantages de chacun de ces types de bases de données, le choix a été

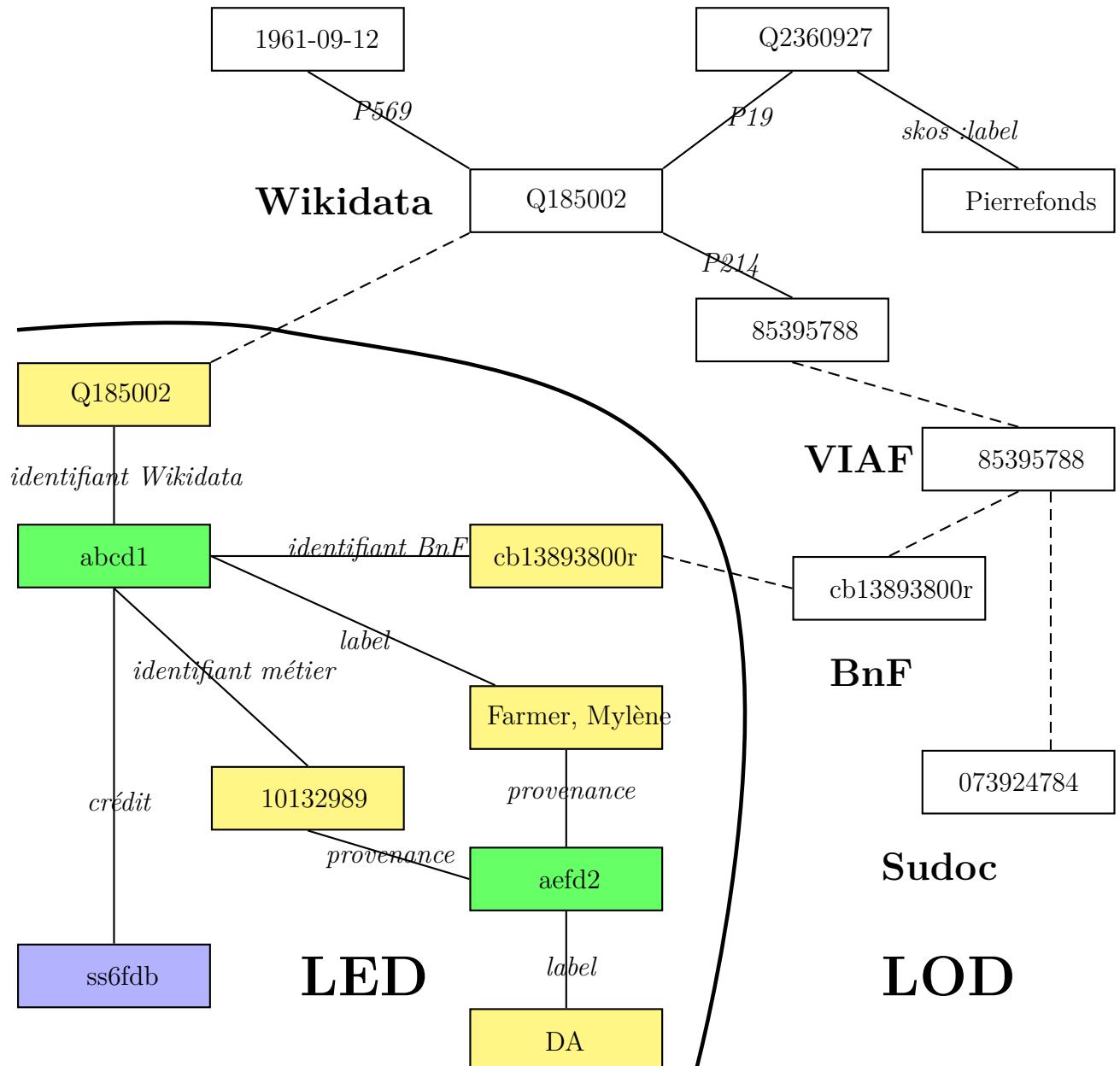


FIGURE 8.4 – Modélisation du concept Mylène FARMER dans le *Lac de données* et le LOD [Données partielles d'exemple. Vert : concept. Jaune : texte. Bleu : instance.]

fait de les utiliser tous les quatre en y dupliquant les données. Ce choix, dans un LED, ne pose pas de difficultés pour la couche supérieure qui est le traitement des données. En effet, les processus ETL — et notamment le logiciel Talend qui permet de les développer — sont créés pour effectuer des extractions de plusieurs bases de données différentes afin de retourner des données transformées, structurées et adaptées aux besoins de la couche supérieure, l'accès aux données pour l'utilisateur. La phase d'accès à ces données n'a alors pas directement accès aux bases sources, mais les obtient par la couche d'abstraction qui est la phase de traitement.

Par le *Lac de données*, l'INA trouve une cohérence dans ses données, les différences existantes entre les diverses bases de données de la DDCOL et de la DJ notamment, ainsi qu'au sein même de la DDCOL avec le DA et la DL, sont effacées au profit d'une individualité de la donnée obtenue par la déconstruction des documents et de l'information. L'effacement de ces différences permet également de normaliser les données autour de mêmes concepts, et ainsi de créer un graphe dans lequel il est possible de circuler entre instances partageant un même concept.

Ces avancées ont été obtenues par le renversement de la pensée du système documentaire et de la place du référentiel. En effet, au lieu de penser la structure des données selon les besoins et les usages finaux — ce qui crée nécessairement autant de structures que de besoins et d'usages —, la réflexion a été retournée pour penser un modèle de données global et unique au sein de l'institution depuis les données elle-mêmes et le contenu intellectuel, depuis les documents conservés à l'INA, eux qui sont l'essence de l'Institut.

8.3 Perspectives d'utilisation

« [L]e rôle central du référentiel va se poursuivre au-delà de [la] réflexion sur l'interopérabilité. En effet, ils sont la pierre angulaire des nouveaux bouleversements autour du *machine learning* et du *deep learning*.¹⁰ »

L'interopérabilité des données — comprises au sens large, avec les métadonnées et les référentiels — est une avancée majeure dans la conception des modèles de données. En effet, en plus de résoudre les nombreuses difficultés qui résultaient de la multiplicité des bases de données à l'INA et de leur conception par les besoins et les usages qui en étaient faits, la centralisation de ces données et leur interopérabilité ouvre des possibilités quant aux nouveaux usages qu'il est possible d'envisager ou de satisfaire.

L'intégration de l'INA dans le *big data* est l'une de ces possibilités : l'utilisation de l'intelligence artificielle permet à la fois une amélioration des descriptions déjà réalisées au

10. *Ibid.*

catalogage, et à la fois la création de nouvelles descriptions. Plus encore que ces actions sur les métadonnées et la création de descriptions de contenu, l'amélioration de la valorisation des documents de l'INA est possible avec le *Lac de données* et l'uniformisation du modèle de données.

8.3.1 Permettre l'intégration des données issues de la description et de la segmentation de vidéos dans le *Lac de données* : réutilisation des concepts et enrichissement des métadonnées

La reconnaissance d'entités nommées est un enjeu essentiel dans la description de documents. Cette dernière est facilitée, depuis quelques années, par des outils nés de programme de recherche sur l'extraction d'entités nommées dans les textes. La classification d'images est également une pratique facilitée par des algorithmes développés par des entreprises comme Google ou Amazon. Cependant, l'extraction d'entités nommées dans des vidéos reste peu pratiquée. L'INA ne dispose alors pas d'outils suffisants et existants pour effectuer une recherche de personne, de logo ou de tableau dans une vidéo. Cette recherche et cette extraction d'entités visuelles dans des vidéos représente l'un des projets de l'INA, DigInPix¹¹.

Dans DigInPix, le rôle du référentiel est essentiel, le référentiel est indispensable au fonctionnement de l'algorithme : le dictionnaire d'entités nommées sur lequel repose l'algorithme permet de reconnaître des logos, des peintures, des personnes physiques ou morales¹². Les bases de données initiales de l'INA n'étant pas suffisamment complètes, les entités ont été enrichies de représentations visuelles trouvées sur le Web, afin d'établir un imposant corpus de comparaison face aux vidéos qui seront à traiter. Une image est tirée de chaque vidéo à intervalle régulier, afin de la comparer à l'ensemble du dictionnaire : plusieurs entités nommées peuvent ainsi être reconnues dans une même image d'une vidéo. De plus, un taux de fiabilité est attribué à chaque rapprochement (Figure 8.5 : Extraction des entités nommées d'un programme de France 2 avec DigInPix).

Avec le projet du *Lac de données*, la création de descriptions de contenus peut aller plus loin encore. En effet, le modèle de données uniifié, comprenant l'ensemble des anciens référentiels de la DDCOL, permet de mettre en relation les données issues automatique-

11. I. national de l'Audiovisuel, *DigInPix - Recherche d'entités visuelles*, DigInPix, URL : <http://diginpix.ina.fr/> (visité le 21/09/2020).

12. Bien que différent par la nature des données stockées, des images, ce dictionnaire est similaire à tout autre dictionnaire comme décrit plus tôt dans notre propos, afin de décrire la diversité d'une entité : « Nous appelons “dictionnaire” une liste d'entités nommées, regroupées pour leur appartenance à certains concepts de niveau hiérarchique supérieur (par exemple, personnes morales, personnes physiques, peintures, bâtiments, etc.). » in Id., *DigInPix - Détails du projet*, URL : <http://recherche.ina.fr/Details-projets/DigInPix> (visité le 21/09/2020)

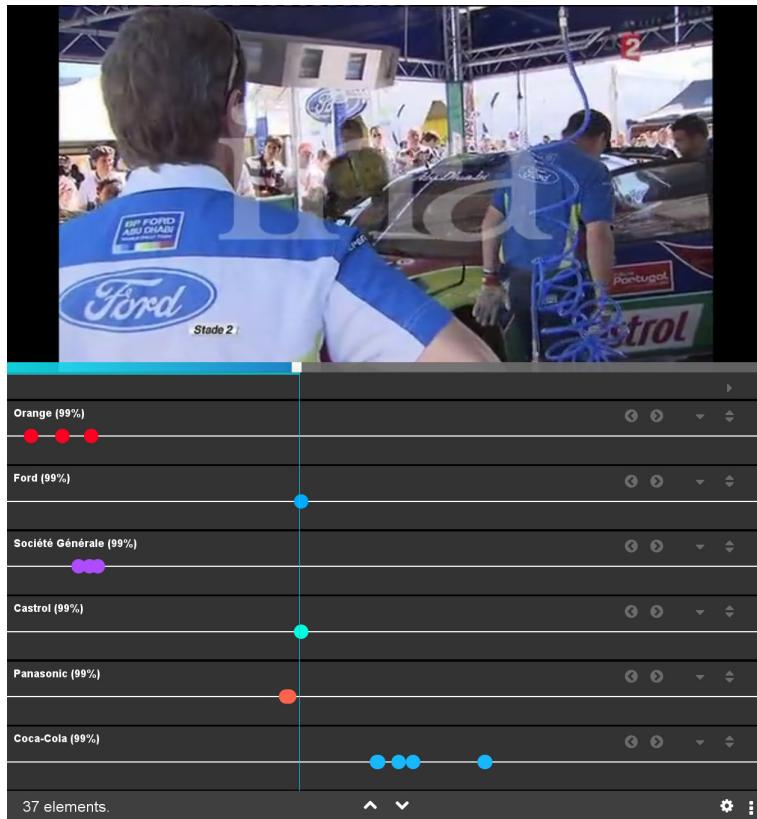


FIGURE 8.5 – Extraction des entités nommées d'un programme de France 2 avec DigInPix
[Source : L'AUDIOVISUEL (Institut national de), *DigInPix - Détails du projet*, URL : <http://recherche.ina.fr/Details-projets/DigInPix> (visité le 21/09/2020)]

ment d'un processus de traitement des vidéos par l'intelligence artificielle avec les concepts du *Lac de données*. La segmentation automatique de vidéos¹³ montre la diversité des utilisations possibles d'un référentiel quand celui-ci est uniforme et centralisé. Ce nouveau projet, au sein de celui du *Lac de données*, débute en 2018 et conduit à la création de multiples outils, tous permettant une description automatique du contenu d'une vidéo — les journaux télévisés et les chaînes d'information. Ainsi, la segmentation automatique par l'intelligence artificielle permet :

- l'établissement d'une grille de programmation à partir des métadonnées fournies par les diffuseurs, les producteurs, ... afin de déterminer les horaires prévus et habituels de chaque programme pour les chaînes d'information
- la classification automatique du programme ou des segments de programme selon une typologie précise — plateau, présentateur, reportage, ...
- la transcription des voix et de la parole, produisant ainsi un texte non formaté à partir duquel une description du contenu est effectuée avec des entités nommées qui en sont extraites et un alignement avec les entités de Wikidata
- l'océrisation des textes présents dans l'image, ce qui produit également un texte qui permet une description intellectuelle du contenu et un alignement des entités

13. Le projet est SAAJ, Segmentation et Analyse Automatique des Journaux télévisés.

nommées avec Wikidata ; cette océrisation concerne notamment les bandeaux des journaux télévisés dans lesquels le nom et la fonction de chaque personne sont indiqués

- la description automatique d'une image par un tagging d'entités nommées
- la reconnaissance de visages afin d'identifier le présentateur du journal télévisé, ou les protagonistes des vidéos
- la reconnaissance d'images et de logos, afin d'enrichir la description déjà précise de la vidéo ou du segment

Chacun de ces outils fonctionne avec ou en relation avec un ou plusieurs référentiels : ils peuvent être internes, c'est à dire propres à l'INA, ou bien externes comme Wikidata qui permet un enrichissement et une ouverture des métadonnées vers l'extérieur. Ces données générées automatiquement sont créées sous le modèle du *Lac de données* et sont par conséquent en relation avec ses concepts.

8.3.2 Faciliter et améliorer le catalogage des documents de l'INA par l'extraction automatique de données

L'apport du projet SAAJ est une description fine et précise de l'ensemble d'une vidéo. Au-delà de la génération automatique de métadonnées dans le *Lac de données*, il devient une aide pour le technicien de gestion des contenus multimédia de l'INA. En effet, il n'a plus à créer les métadonnées associées au document, mais à superviser leur qualité et leur véracité : « Le documentaliste « humain » est-il destiné à passer du statut de producteur de données à celui de contrôleur de la qualité des fruits de l'automatisation ? »¹⁴. Cet aspect de contrôle qualité est un usage indirect des données générées automatiquement : il est positif pour la création précise, fine et complète de métadonnées sur un programme ; mais il contraint à un changement de pratiques de catalogage, où l'humain n'a plus le rôle principal, qui est intellectuel, dans lequel il décrit le document et son contenu.

Cette amélioration et cette facilitation du travail de catalogage a lieu depuis des données nouvelles générées par l'intelligence artificielle. Cependant, le contrôle de la qualité des métadonnées, et leur enrichissement, peut également passer par un traitement *a posteriori*. En effet, l'apparition du Web sémantique et son adoption par un grand nombre d'institutions a poussé l'INA, associé à d'autres institutions, à mener le projet Qualinca entre 2012 et 2015 afin « d'améliorer la richesse, la cohérence et l'interopérabilité des métadonnées du système documentaire de l'Ina à travers la mise en œuvre d'une activité de recherche dans le domaine des techniques de liage de données »¹⁵. Qualinca repose sur de nombreux enjeux, comme la possibilité de partager des identifiants communs entre

14. E. Alquier, J. Carrive et S. Lalande, “Production documentaire et usages”..., p.134.

15. *Ibid.*, p.129.

les différents métiers, l'amélioration des descriptions de contenus grâce aux données extérieures du LOD, mais également d'effacer les ambiguïtés des termes des lexiques de l'INA.

Se basant sur deux algorithmes, ProbFr et Agreg, Qualinca s'est surtout tourné vers les alignements de corpus de musique, et d'homonymes de personnes physiques et d'émissions. Dans cet alignement des homonymes avec le LOD, la base DBpedia — Wikidata n'est né qu'en 2014 —, les résultats sont peu exploitables et se heurtent, comme nous avons pu le constater lors de l'alignement des personnes physiques avec Wikidata, au langage naturel des fonctions que les algorithmes sont incapables de dépasser : sur 5000 Jacques MARTIN, 667 différents ont été identifiés par les algorithmes¹⁶.

L'extraction automatique d'entités nommées a permis la création de nouvelles métadonnées associées non pas au matériel ou aux données de diffusion, mais au contenu intellectuel des vidéos, ainsi que l'apport d'une aide au catalogage par la qualité des entités fournies et leur précision.

8.3.3 Améliorer la valorisation des documents et offrir une meilleure expérience utilisateur

La centralisation des données de l'INA au sein du *Lac de données* ouvre des possibilités pour la valorisation des documents auprès de tous les publics¹⁷. D'abord, cette centralisation permet la création de multiples applications pour l'utilisateur, sans que cela ne modifie la structure des données¹⁸ ; ainsi, la présence de l'INA en est modifiée par l'apparition d'un *hub* regroupant l'ensemble de l'offre numérique de l'Institut¹⁹.

Cette centralisation de l'offre est aussi présente pour les usages internes avec la création d'une nouvelle interface de consultation des métadonnées et des documents en eux-mêmes, Notilus, afin d'éviter la consultation croisée de multiples interfaces selon la provenance du document comme cela était le cas avant le *Lac de données*. Par cette interface, le DL, le DA, puis la DJ et l'ensemble des professionnels de l'INA, ont accès aux mêmes données et aux mêmes documents, en un point unique, une interface de consultation qui reconstitue les instances du *Lac de données*.

16. *Ibid.*, p.133.

17. Pour l'ensemble de l'offre disponible, voir la Chapitre 3 : Les référentiels à l'INA.

18. Voir Annexe K : Repenser l'infrastructure (Figure K.1 : Schéma du *Lac de données* depuis le stockage jusqu'à l'accès aux données).

19. « L'autre grand défi posé à l'Institut, c'est celui de l'accessibilité de ses propositions. Les rassembler au sein d'un grand portail numérique, un hub qui offrira en quelques clics un accès renouvelé, simplifié et cohérent à l'ensemble des activités, contenus et services de l'Ina, est ainsi l'objectif qui mobilise aujourd'hui toute l'entreprise à l'horizon de 2019. » in Laurent Vallet, *Ina - L'éditorial du président de l'Institut national de l'audiovisuel*, URL : <http://institut.ina.fr/institut/organisation/l-editorial-du-president> (visité le 21/09/2020)

L'amélioration de l'expérience utilisateur est également une priorité dans une période où la modification des pratiques est radicale : ces pratiques sont quasiment toutes numériques et contraignent l'INA à s'adapter. Si le site <https://www.ina.fr> est né dès 2009, le *Lac de données* va pouvoir lui apporter d'importantes améliorations. En effet, la présence des référentiels y est limitée et limite les possibilités de rebonds de la part de l'utilisateur²⁰. Ainsi, les liens des personnes au générique ne renvoient pas à une vedette personne de l'INA, mais à des résultats de recherche sur le nom de cette personne.

The screenshot shows a video page from ina.fr. At the top, it says "Mimi Mathy et Yves Lecoq jouant à l'école des Fans". Below that, there's a timestamp ("video | 05 déc. 1992 | 1057 vues | 00min 27s"). The main content area has a summary: "Sur le plateau du Téléthon 92, Yves LECOQ imite tout à coup Jacques MARTIN dans 'L'école des fans', adressant à Mimi MATHY comme à une enfant.". To the right, under "Production", it lists "Producteur ou co-producteur: Association Française contre la Myopathie, JACQUES MARTIN PRODUCTION, France". Below this, under "Générique", it lists "Réalisateur: Jean Pierre Spiero", "Participant: Mimi Mathy, Yves Lecoq", and "Présentateur: Gérard Holtz". On the far right, there's a sidebar titled "PLUS DE CONTENUS SUR" with various tags: solidarité, recherche médicale, maladie génétique, diffusion en direct, Martin Jacques, AFM, handicapé, myopathie, don-solidarité, campagne de solidarité, and émission télévisée.

FIGURE 8.6 – Métadonnées associées à un document sur ina.fr [Source : <https://www.ina.fr/video/I11297765/mimi-mathy-et-yves-lecoq-jouant-a-l-ecole-des-fans-video.html>]

Les possibilités offertes par le *Lac de données* sont multiples et non pouvons en imaginer certaines, basées sur le seul usage des concepts et de leurs relations, qui faciliteraient la recherche de l'utilisateur. Ainsi, de même que la BnF, la création de vedettes de personnes est envisageable afin de regrouper en une même page les informations biographiques, ainsi que les documents liés (les instances) ou bien les thématiques principales (les concepts liés). Les termes d'indexation et de description des vidéos peuvent également être concernés par ce regroupement d'informations et de liens. Cependant, plus encore que ces regroupement de métadonnées, d'instances et de concepts relatifs à un concept, il est désormais possible, avec les liens établis avec le LOD, d'obtenir des informations manquantes et d'enrichir les données proposées à l'utilisateur : un lien peut être inséré, comme c'est le cas dans VIAF, ou bien les champs peuvent être directement remplis sur la page HTML.

Ces possibilités ne sont possibles que grâce à la déconstruction de l'information dans le *Lac de données*, permettant alors une grande modularité des données dans les usages qui en sont faits. Ces usages ne sont pas tous nés et le *Lac de données* doit pouvoir permettre à l'INA de les remplir sans avoir recours à un modification du modèle de données ou à la création d'une nouvelle de données. Ainsi, s'il est nécessaire de publier les données²¹

20. Voir Figure 8.6 : Métadonnées associées à un document sur ina.fr.

21. Cet aspect semble difficile pour l'INA en raison des données personnelles qui y sont conservées.

sur le Web et plus particulièrement sur le Web de données, une représentation RDF est possible ; avec la présence forte de l'INA sur les réseaux sociaux, il peut être envisager de créer des publications automatiquement à partir de tags issus de concepts ; ...

Le référentiel a atteint une place centrale dans le *Lac de données* : l'ensemble des applications et des sites de l'Institut fonctionnent ou vont fonctionner depuis ce silo de métadonnées qui a été pensé selon la donnée et non plus les besoins. Ces besoins, évolutifs et dépendants de la période, ne peuvent pas tous être prédits, ce qui a conduit à la constitution d'un modèle de données souples et d'un processus intermédiaire de traitement de ces données de manière à offrir à chaque application les données qui lui sont nécessaires.

Cependant, la loi de 2016 pour une République numérique (*LOI n° 2016-1321 du 7 octobre 2016 pour une République numérique (1)* - Légifrance, 7 oct. 2016, URL : <https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000033202746/> (visité le 21/09/2020)) encourage à la publication des données de référence qui peuvent être réutilisées par d'autres services ou d'autres institutions.

Chapitre 9

Centraliser les référentiels de l'INA dans le *Lac de données* : l'exemple de l'alignement de deux référentiels de personnes physiques

La DDCOL possède un unique référentiel de personnes physiques comme nous avons pu le constater plus tôt¹. Le *Lac de données* étant un LED, il tire son intérêt de la mise en commun de différentes bases de données et jeux de données. Ainsi, c'est l'ensemble des données et des métadonnées de l'INA qui entrent dans ce *Lac* : la base de données de la DJ fait alors l'objet de ce processus de migration vers le *Lac de données*. Cependant, bien que la DJ représente un silo de données distinct de celui de la DDCOL, ils partagent tout de même certaines caractéristiques comme l'utilisation massive de personnes physiques, qui sont réunies dans un référentiel dans chaque service.

La migration des données de la DDCOL s'achève en fin d'année 2020 et laisse la place à celle de la DJ : afin d'éviter toute redondance de concepts dans le *Lac de données*, il convient de rechercher pour une personne physique de la DJ son équivalent dans le *Lac de données* — par conséquent dans l'ancien référentiel des personnes physiques et morales de la DDCOL qui a été déconstruit sous forme de concepts². L'exécution de cet alignement montre à lui seul les problématiques liées au langage naturel, ainsi que la nécessaire présence humaine qui doit superviser les résultats issus d'une automatisation de tâches.

1. Voir section 2.3 : Passer du texte libre à un vocabulaire contrôlé : aligner des notes qualité et un thésaurus de noms communs et ?? : ??.

2. Voir Chapitre 8 : Le *Lac de données* de l'INA : le référentiel au centre du modèle.

9.1 Des jeux de données différents en de multiples points

L’alignement des personnes physiques de la DDCOL avec Wikidata avait déjà démontré l’importance de la donnée structurée afin de créer des points de contacts entre les deux jeux de données et procéder à leur alignement. Les problématiques liées à la graphie et aux différences de structure ont aussi compliqué cet alignement. Dans le cadre de la mise en relation entre le référentiel des personnes physiques de la DJ avec celui de la DDCOL, ces points de contact sont réduits au minimum et peuvent interroger quant à la possibilité de réaliser des alignements sûrs, ou les plus sûrs possibles.

9.1.1 Enjeux

Le *Lac de données* n’étant pas conçu à partir des besoins métier mais des données, le référentiel des personnes physiques de la DJ, se présentant sous la forme d’une table *PERSONNE* de la base de données, ne peut pas être conservé dans sa structure actuelle. En effet, il est uniquement adapté aux besoins de la DJ et ne correspond pas aux usages que pourrait en faire la DDCOL ou l’utilisateur final des applications de lINA. Dans le but d’intégrer ce référentiel dans le *Lac*, il est nécessaire de l’aligner avec les concepts existants, issus du référentiel des personnes physiques de la DDCOL. Ainsi, un double enrichissement a lieu, celui des concepts par les données de la DJ, et celui de la DJ par les données des concepts. Cependant, cet enrichissement devient invisible dans le *Lac de données* puisqu’il n’y a plus de notion de référentiel, et que les distinctions entre DJ et DDCOL sont volontairement effacées au profit d’une structure de données plus souple.

Cet alignement a pour finalité l’ajout d’un lien *provenance – DJ* à un concept quand le matricule de la DJ et le concept sont identiques, ou bien la détection des nombreuses personnes de la DJ qui ne sont pas des concepts. En effet, la DJ ayant pour fonction de repérer les ayants-droits et ouvrants-droits de personnes liées à un extrait audiovisuel, ces ayants-droits et ouvrants-droits ne sont par conséquent pas spécifiquement dans la description des documents audiovisuels réalisée par la DDCOL car ils n’interviennent à aucun moment dans ces documents. Ainsi, nombreuses sont les personnes de la DJ qui n’ont pas d’équivalent dans la DDCOL et qu’il est nécessaire de repérer afin de leur créer *in fine* un concept dans le *Lac de données*.

Cette différence entre les jeux de données montre une nouvelle fois comment se sont formées les bases de données — selon les usages et les besoins — qui sont à migrer dans le *Lac de données* : à cause de cette différence, il devient difficile d’estimer l’efficacité et le rendement du processus d’alignement qui va être réalisé. En effet, connaître la raison

de l'absence d'alignement de certains matricules des personnes de la DJ sera uniquement possible par une action humaine. Face à ces enjeux et aux problématiques soulevées par le seul historique des bases de données, l'alignement des deux référentiels comporte plusieurs difficultés supplémentaires déjà évoquées dans les chapitres précédents.

9.1.2 Points de contact

Trouver des points de contact entre deux jeux de données, plus encore entre deux référentiels, est essentiel lors d'un alignement : plus ces points de contact sont nombreux, plus les comparaisons sont nombreuses et les alignements sûrs. Cependant, les référentiels de la DDCOL et la DJ n'en partagent que peu — sept. De plus, ces points de contact nécessitent la présence de l'information de chaque côté, ce qui est peu le cas entre la DJ et la DDCOL.

Dans le *Lac de données*, les concepts disposent notamment d'attributs indiquant le nom, le sexe, les dates de naissance et de décès, ainsi que la note qualité. Cette note qualité n'étant pas scindée dans le *Lac de données*, il est nécessaire dans cet alignement d'en extraire la fonction, ou les fonctions, de la personne, en supprimant la mention des pays d'exercice.

Si la DJ dispose de beaucoup de données personnelles pour mener à bien ses missions, les données permettant un alignement documentaire sont plus restreintes : hormis le nom et le sexe, seule une date de décès est disponible, ainsi qu'une contribution. En effet, seule la date de décès intéresse le service juridique pour ses applications dans le droit et le reversement des droits aux ayants-droits ou ouvrants-droits : conserver une date de naissance n'a par conséquent aucun usage dans la DJ.

Le référentiel des personnes de la DJ présente une petite normalisation avec les contributions : celles-ci ne sont pas du texte libre, mais du texte contrôlé et choisi parmi une liste d'une vingtaine d'entrées. Ce contrôle du vocabulaire permet dans l'alignement des rendements meilleurs après, nous le verrons, un traitement préalable des données des notes qualité de la DDCOL.

Cependant, les points de contact identifiés pour la DJ et la DDCOL sont peu nombreux, ce qui complique la détection des homonymes et diminue la fiabilité des alignements qui seront réalisés.

9.1.3 Divergences

En plus de ces difficultés sur la quantité des points de contact, les deux référentiels diffèrent par leurs structures et leurs graphies. D'abord, les niveaux de description des

mêmes attributs sont différents. En effet, alors que le nom du concept du *Lac de données* se compose de la forme *Nom, Prénom*, l'état-civil stocké à la DJ est divisé en deux attributs : un nom, et un prénom. Ainsi, avant d'effectuer l'alignement, il est nécessaire de scinder le nom du concept afin de récupérer le nom et le prénom séparément.

Le jeu de données de la DJ offre également deux autres attributs, les pseudos de nom et de prénom de chaque personne. Afin d'utiliser ces deux attributs supplémentaires, il est nécessaire de leur trouver un point de contact dans les concepts issus de la DDCOL : ainsi, il a été considéré qu'un nom de concept ne possédant pas de virgule est un pseudo. Par conséquent, ce pseudo issu des concepts peut être comparé avec le pseudo du nom de la DJ³.

En plus de ces différences de niveaux de description, les graphies ne sont pas les mêmes. D'abord, les données de la DJ sont en majuscules, alors que celles de la DDCOL sont en minuscules. Si cette difficulté n'est pas majeure, elle nécessite tout de même un traitement dans l'ETL avant de pouvoir procéder à un alignement. De même, afin d'éviter toute variation dans des chaînes de caractères renvoyant à une même personne mais aux graphies différentes, les accents et la ponctuation sont retirés. Les dates, de même que lors des alignements décrits dans les chapitres précédents, sont réduites à la seule année. La difficulté majeure posée dans l'alignement de deux référentiels de personnes est la graphie et l'utilisation des particules des noms. En effet, l'utilisation des particules n'est pas normalisée dans l'Institut, ce qui conduit à la présence de Louis de FUNÈS dans la DDCOL, alors que la DJ conserve la forme Louis DE FUNÈS. Les pratiques d'écriture des noms à particules étant constantes à la DDCOL, il est possible de transférer cette particule⁴ dans le nom afin d'obtenir Louis DE FUNÈS dans chaque jeu de données.

Enfin, afin de donner aux alignement une fiabilité plus importante, il est essentiel de prendre en compte le texte des notes qualité pour le comparer avec les contributions de la DJ. Les notes qualité de la DDCOL comportant plus de vingt mille fonctions différentes, il n'est pas possible de les faire correspondre chacune avec l'une des contributions de la DJ. Pour cela, seules les contributions et les fonctions des notes qualité les plus courantes ont fait l'objet d'un alignement manuel pour faciliter l'alignement automatique qui va suivre ; cinq de ces contributions ont ainsi été pu être traité :

- le terme *Journalisme* de la DJ est remplacé par « journaliste » ;
- *Artiste interprète* est remplacé par « chanteu »⁵

3. Dans les données de la DJ, c'est le pseudo-nom qui comporte le pseudonyme courant d'une personne ; l'attribut pseudo prénom n'est utilisé que pour indiquer une variante du prénom de cette personne.

4. Cette particule n'est pas exclusivement *de*, elle peut être de l'une des formes suivantes : *de, des, du, de la*

5. Les terminaisons sont enlevées dans ces termes de remplacement afin de prendre en compte les variantes de graphie liées au pluriel et au féminin que l'on peut trouver dans les fonctions de la DDCOL.

- « realisat » remplace *Réalisation*
- « composit » remplace *Composition musicale*
- enfin, *Réalisation associée* est remplacé par « realisat »

Les chanteurs, les compositeurs, les réalisateurs et les journalistes étant les fonctions les plus courantes dans les deux jeux de données, elles ont été repérées puis traitées. Cependant, une majorité de fonctions ne pourra pas être alignée avec les contributions de la DJ, et par conséquent limitera la confiance accordée aux alignements.

La centralisation de référentiels et de données est nécessaire pour les systèmes documentaires, mais la reprise de ces référentiels et de ces données peut être compliquée par les structures et les normes divergentes selon les jeux de données. Cette absence d'uniformisation annonce déjà des résultats faibles et limités dans la confiance que l'on peut leur accorder. Dans le cas de l'alignement des référentiels de personnes physiques de la DJ et de la DDCOL à l'INA, les points de contacts sont peu nombreux et peu spécifiques⁶.

DJ	DDCOL
Nom	Nom
Prénom	Prénom
Pseudo nom	Pseudo
Pseudo prénom	
Sexe	Sexe
Date de naissance	Date de naissance
Contribution	Fonction

TABLE 9.1 – Points de contact entre les référentiels de la DJ et de la DDCOL

9.2 Établir une méthodologie particulière d'alignement

En raison des difficultés identifiées dans la section 9.1 : Des jeux de données différents en de multiples points, un alignement simple, n'apportant aucune indication de fiabilité, n'est pas possible. De même, les alignements réalisés dans les chapitres précédents utilisent chacun les jeux de données initiaux jusqu'à la fin du traitement, sans en retirer au fur et à mesure les concepts qui viennent d'être alignés. L'alignement des référentiels de la DDCOL et de la DJ se distingue des précédents par la nécessité d'une méthodologie particulière, basée sur un indice de confiance attribué à chaque alignement, et sur une succession d'étapes, représentant les niveaux de confiance apportés au type de jointure utilisé.

6. Voir Table 9.1 : Points de contact entre les référentiels de la DJ et de la DDCOL.

9.2.1 Créer un indice de confiance pour chaque alignement

Les points de contact entre les jeux de données n'ont pas tous la même valeur dans un alignement. En effet, l'état civil d'une personne, bien qu'essentiel dans un alignement, peut conduire à aligner deux homonymes : c'est pourquoi les points de contact comme les noms, les prénoms ou les pseudos peuvent être considérés comme ayant une faible valeur dans le processus d'alignement. De plus, leur octroyer une valeur forte conduirait à surévaluer les alignements réalisés sur la simple comparaison des noms et prénoms sans autre point de comparaison par rapport aux alignements qui n'auront pas été possibles.

En revanche, la valeur des points de comparaison significatifs est considérée comme forte : il s'agit de la date de décès ou de la contribution. En effet, la probabilité que les états civils et les dates de décès de deux homonymes soient identiques est très faible, ce qui peut permettre de donner à cet alignement une valeur plus forte. De même, la correspondance entre une contribution et une fonction est considérée comme très fiable quand les états civils ont déjà été rapprochés : pour cette raison, le point de comparaison sur la contribution est celui qui possède l'indice de confiance le plus élevé, puisqu'il est le point le plus spécifique, et le plus difficile à faire correspondre.

Enfin, la comparaison du sexe permet également d'augmenter la fiabilité d'un alignement. Dans la majorité des alignements qui sont réalisés, la comparaison peut sembler évidente à l'humain, mais dans certains cas, comme pour les prénoms *Dominique*, elle est nécessaire et permet la conservation ou non de l'alignement.

L'indice de confiance permet une priorisation des points de comparaison, une hiérarchisation de ces derniers. Il se forme à partir de la somme des scores de chaque point de comparaison. Ainsi, dans le cas de l'alignement des référentiels de personnes de la DJ et de la DDCOL, l'indice de confiance varie entre 0 — quand l'alignement n'a pas pu être réalisé — et 9 — quand tous les points de comparaison ont été réalisés avec succès — selon les scores de la Table 9.2 : Scores attribués à chaque point de comparaison.

Point de comparaison	Score
Nom	1
Prénom	1
Pseudo nom	1
Pseudo prénom	1
Sexe	1
Date de décès	2
Contribution	2

TABLE 9.2 – Scores attribués à chaque point de comparaison

9.2.2 Des étapes exclusives

L'attribution d'un indice de confiance ne permet de résoudre que la problématique de l'évaluation finale des alignements. Il subsiste néanmoins une seconde problématique, celle de la présence dans les alignements réalisés de doublons, c'est à dire de personnes de la DJ alignés avec plusieurs concepts. Une solution pourrait être de supprimer les alignements présents dans ce cas. Or, ce cas survient fréquemment : supprimer les alignements concernés réduirait la quantité de résultats finaux.

Ainsi, après une première priorisation des points de comparaison, il est nécessaire d'effectuer ensuite une priorisation des combinaisons de ces points de comparaison. Quatre étapes principales ont été identifiées et constituent cette priorisation.

D'abord, les alignements réalisés avec le nom, le prénom, les pseudos et la correspondance des fonctions sont considérés comme ceux étant les plus sûrs pour effectuer les rapprochements entre les concepts de la DDCOL et les matricules de la DJ. Ces alignements, comme ceux des étapes suivantes, sont des jointures⁷ entre les deux jeux de données réalisées dans l'ETL Talend avec le composant associé, un tMap (Figure 9.1 : L'alignement des personnes de la DJ et de la DDCOL pour une jointure dans un tMap de Talend.).

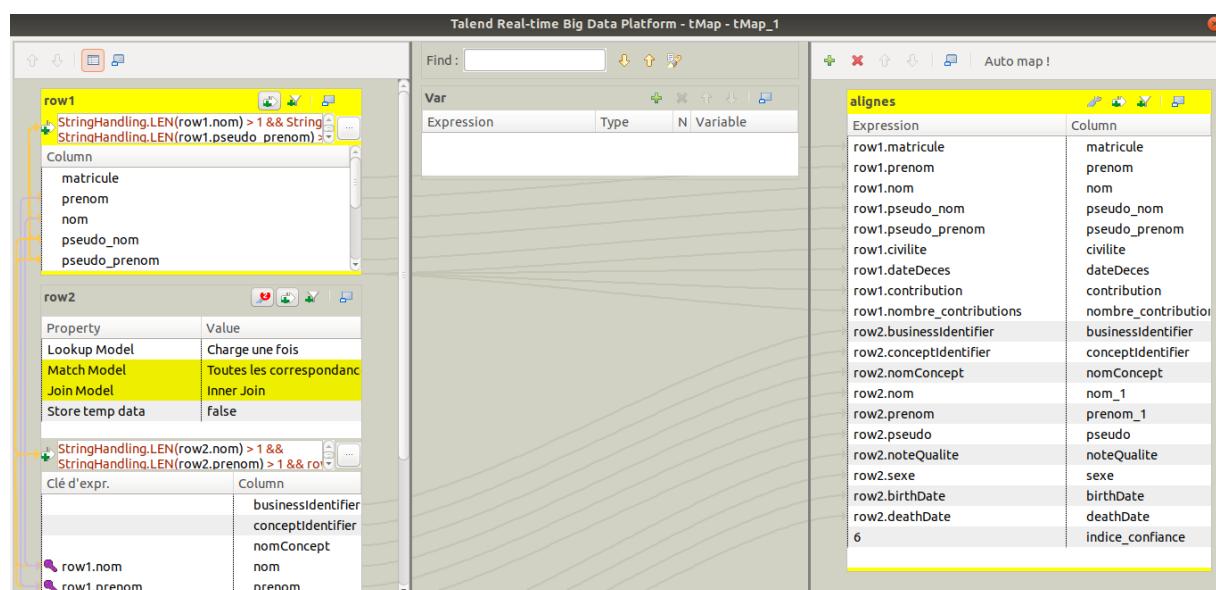


FIGURE 9.1 – L'alignement des personnes de la DJ et de la DDCOL pour une jointure dans un tMap de Talend.

Ensuite, les alignements effectués avec le nom, le prénom et les pseudos, sans avoir pu faire correspondre les fonctions, sont la seconde étape.

La troisième étape, comme la quatrième, tente de rapprocher deux personnes en prenant en compte les différences de graphie qui peuvent exister. Ainsi, les alignements sont réalisés

7. Afin de ne récupérer que les alignements qui ont été réalisés, ces jointures sont de type *inner join*.

sur les pseudos, et sur le prénom de la DJ commençant par le prénom de la DDCOL⁸. Cette étape permet l'alignement d'une même personne ayant à la DDCOL le prénom *Louis* et à la DJ le prénom *Louis Marie*.

Enfin, l'ensemble des combinaisons possibles étant couvert, il est nécessaire d'effectuer une dernière étape pour effectuer non pas des alignements fiables — ce qui est le but des trois premières étapes — mais des alignements permettant d'apporter une aide à un opérateur humain en proposant plusieurs concepts qu'il est possible d'aligner avec un matricule. Ce rapprochement particulier, réalisé sur les seuls nom et prénom, autorise par conséquent la présence de plusieurs concepts alignés avec un même matricule, notamment dans le cas d'homonymes.

Distinguer ces quatre étapes permet, à l'issue de chacune d'elles, de récupérer ce qui n'a pas été aligné, tant du côté de la DJ et de la DDCOL, afin d'effectuer l'étape suivante avec uniquement ces données non alignées (Figure 9.2 : L'orchestration de la première étape dans l'ETL Talend.). Cette récupération évite de créer des alignements doubles avec des concepts différents entre les étapes. C'est également à l'issue de cette récupération que les résultats des jointures précédentes sont comparés afin de supprimer les matricules alignés plusieurs fois, et d'attribuer les scores pour le sexe et la date.

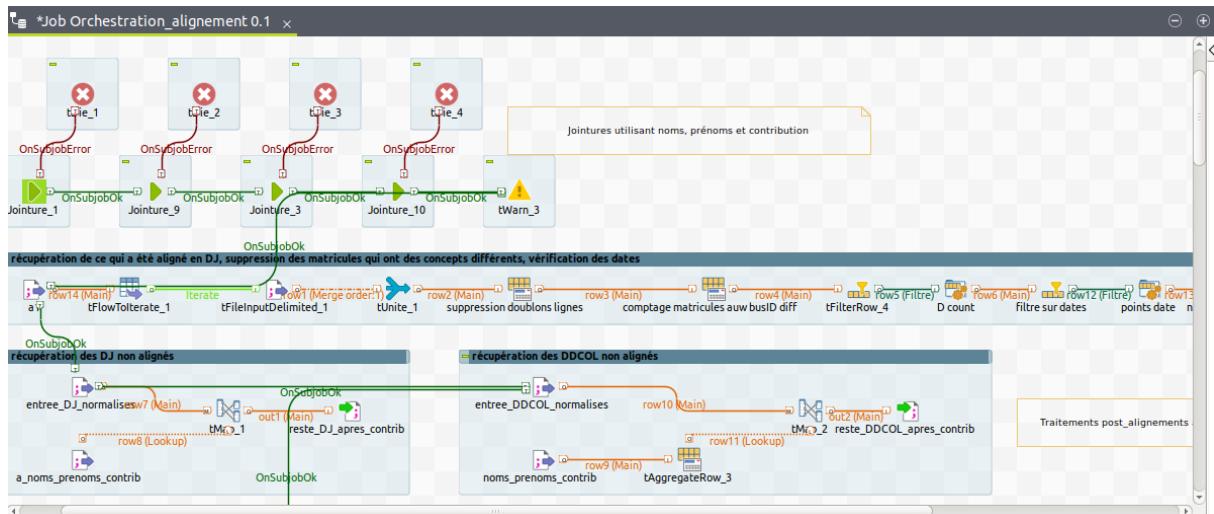


FIGURE 9.2 – L'orchestration de la première étape dans l'ETL Talend.

Cependant, la limite de la récupération des matricules des personnes non alignées de la DJ et de la DDCOL est dans la priorisation qui a été faite des étapes. En effet, elle se base uniquement sur des critères définis en amont par un humain selon l'observation des cas généraux d'alignements : bien que ce soit une méthode apportant de la fiabilité aux alignements, cette fiabilité n'est pas nécessairement celle qui est la plus optimale. De

8. Une exception a été créée dans cette étape pour les prénoms *Jean* et *Anne*.

même, cette récupération pourrait avoir lieu après chaque alignement de matricule : cela ajouterait néanmoins une nouvelle priorisation, basée sur l'ordre de passage, qui est plus arbitraire encore, puisque cet ordre de passage des matricules de la DJ dans le processus d'alignement n'est régi par aucun ordre significatif⁹.

Face aux nombreuses difficultés, l'indice de confiance et la priorisation des étapes permettent de réduire les mauvais alignements, et d'apporter une précision sur la fiabilité d'un alignement. Cependant, l'automatisation de ce processus présente des limites comme la définition de règles dirigeant le processus d'alignement. L'alignement des soixante-dix mille matricules de la DJ et des plus de trois cent mille concepts de la DDCOL ne peut pas être réalisé sur la base de quelques règles puis être considéré comme fiable.

9.3 Des résultats à la hauteur des données initiales

Les résultats de l'automatisation d'un processus de traitement de données dépendent entièrement de la qualité des données initiales. Les difficultés propres aux différences de structures entre les jeux de données ou aux différences de graphie, additionnées à celles posées par la volonté d'avoir des alignements fiables et sans doublons, sont autant de facteurs qui réduisent l'efficacité de l'alignement automatique de deux référentiels.

Dans le cas de l'alignement des référentiels de personnes de la DJ et de la DDCOL, les résultats reflètent les difficultés rencontrées, tant dans les quantités de résultats que dans les indices de confiance attribués. Cette hétérogénéité des résultats conduit à la nécessité d'une supervision humaine des alignements réalisés et non réalisés.

9.3.1 Des résultats hétérogènes reflétant les multiples difficultés

Environ soixante pour cent des matricules de personnes de la DJ ont trouvé leur équivalent dans la DDCOL. Ce résultat, bien que faible, reflète les difficultés rencontrées, ainsi que la spécificité des usages de chaque référentiel. En effet, la DJ utilise les personnes pour leur verser des droits, ce qui signifie alors que ces personnes ne sont pas spécialement des acteurs des documents conservés à la DDCOL pour lesquels le référentiel des personnes physiques a été créé. Il est par conséquent normal de ne pas pouvoir aligner tous les matricules de la DJ avec la DDCOL, cette dernière n'ayant pas besoin de conserver les ayants-droits de chaque personne du référentiel.

La répartition des indices de confiance (Figure 9.3 : Répartition des indices de confiance après les alignements entre les référentiels de personnes de la DJ et de la DDCOL) reflète quant à elle à la fois la qualité des données initiales, ainsi que le processus

9. Une base de données relationnelle n'étant pas ordonnée.

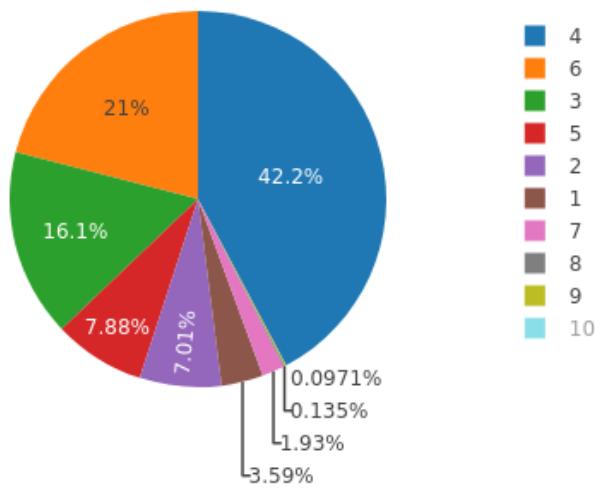


FIGURE 9.3 – Répartition des indices de confiance après les alignements entre les référentiels de personnes de la DJ et de la DDCOL

d'alignement en lui-même. En effet, les différences de niveaux de précision dans les jeux de données initiaux conduisent à l'impossibilité d'utiliser certains points de comparaison, ce qui réduit l'indice de confiance : ces différences peuvent être une absence de données d'un côté de l'alignement, ou bien une divergence de graphie ou de structure qui n'a pas pu être corrigée lors du traitement préalable des données. Ainsi, l'indice 4 est celui le plus présent en raison du faible nombre de points de comparaison qui le permettent : le nom, le prénom ainsi que la date ou une fonction suffisent à attribuer cet indice.

De plus, l'enchaînement des étapes entraîne également une diminution des indices de confiance : les étapes considérées comme prioritaires sont aussi celles qui utilisent les points de comparaison à la plus forte valeur. Ainsi, les indices de confiance attribués dans les alignements sont globalement inférieurs à cinq, et peu peuvent être considérés comme fiables quand l'indice est supérieur à cinq ou six.

9.3.2 Une supervision humaine nécessaire

L'incertitude entourant la majorité des alignements conduit, comme cela est le cas pour le catalogage après la génération automatique de données, à utiliser une supervision humain. En effet, seule l'expertise et la réflexion d'un agent humain peut, ou non, confirmer les alignements produits. Seulement, cet agent humain va se heurter également à certaines problématiques posées dans l'automatisation : l'absence d'informations dans un matricule de la DJ ou dans un concept du *Lac de données* ne permettra pas d'affirmer si les deux personnes alignées automatiquement sont réellement identiques et peuvent être associées.

Outre ce rejet ou cette acceptation des alignements réalisés automatiquement, la

supervision humaine doit pouvoir modifier ce qui lui est proposé, ou bien pouvoir créer de nouveaux alignements. En effet, les jointures effectuées dans Talend¹⁰ ne prennent pas en compte toutes les possibilités des deux jeux de données¹¹ pour des raisons de fiabilité de ces possibilités dans le processus d'alignement. L'agent humain est seul capable de rechercher dans les données de la DDCOL un concept selon des critères que l'intelligence humaine offre : ils peuvent être l'inversion de deux lettres suite à une coquille dans la graphie, la francisation ou la traduction d'un terme étranger, l'inversion de deux prénoms, la connaissance d'un autre pseudonyme de la personne qui n'est pas spécifié dans les données de la DJ, ...

L'action humaine est toujours nécessaire, même avec un processus automatique d'alignement entre des données. Cette action est essentielle pour obtenir des données cohérentes et fiables dans un nouveau modèle de données. En effet, ces données étaient initialement cohérentes et sûres dans leurs bases de données respectives : elles doivent retrouver cette cohérence et cette fiabilité, même après un traitement automatique. Pour cela, la supervision humaine est nécessaire pour s'en assurer et proposer le cas échéant des solutions.

Dans le projet de centralisation des données de l'INA, la centralisation des référentiels est indispensable, ces derniers étant devenus les pivots du système d'information. La conservation d'un référentiel utilisé par un seul jeu de données, celui de la DJ, ne correspond pas aux principes du LED et du *Lac de données*. Ainsi, sa migration dans ce *Lac* a nécessité un traitement important pour aboutir à l'alignement de ses matricules avec les concepts du *Lac de données*. Face aux résultats de cet alignement, l'indice de confiance attribué permet une meilleure visibilité du travail effectué automatiquement, et permet aux superviseurs, qui sont nécessaires, d'approuver ou de refuser chaque alignement, afin d'éviter l'introduction d'erreurs dans le *Lac de données*.

10. Elles sont au nombre de onze.

11. Il existe quarante-deux (sept attributs sont disponibles dans la DJ, et six dans la DDCOL) jointures de stricte égalité possibles.

Avec la théorisation du labyrinthe à la fin de l'Ancien Régime et du modèle-réseau par Umberto ECO, les référentiels ont trouvé une nouvelle place dans les systèmes d'information. En effet, bien que déconstruits en noeuds et en liens, ils occupent désormais une place centrale dans ces systèmes. Ils sont la cible de toutes les attentions afin de pouvoir s'adapter aux besoins de l'intelligence artificielle, aux usages nouveaux de l'utilisateur sur le Web, aux besoins des professionnels de l'information et de la documentation. Les jeux de données peuvent devenir des jeux de données de référence dès lors qu'il sont réutilisés ou partagés.

Seulement, la centralisation des données décentralisées par le modèle-réseau est apparue comme nécessaire et a créé des hubs de liens qui permettent d'accéder en un lieu unique à un corpus de liens relatif à un seul concept. Le parcours de graphe et de liens ne requiert alors qu'un simple passage par ce hub pour trouver la ressource recherchée. Au sein d'un LED, il est également possible de créer un hub de liens : c'est alors le concept lui-même qui devient un hub grâce aux identifiants qu'il peut porter.

L'ouverture d'une institution vers le Web de données a été facilitée par le Web de données et le modèle-réseau. Cependant, le but premier de la création d'un LED est la mise en cohérence des données propres à une institution, ce qui a conduit à la création du *Lac de données* à l'INA. Or, cette mise en cohérence, des référentiels notamment, est compliquée par les pratiques de rédaction des termes et de choix des attributs stockés selon les usages qui les régissent.

Conclusion

Ce mémoire s'est attaché à montrer l'évolution de la structure et de la place des référentiels dans les systèmes documentaires des institutions patrimoniales. L'exemple du *Lac de données* de l'INA a permis d'illustrer les structures des référentiels, les problématiques associées, ainsi que les solutions adoptées afin d'obtenir une place nouvelle pour ces référentiels dans un modèle de données repensé.

La recherche du moyen de classement de documents le plus efficace, ainsi que de l'outil de description le plus adapté, est une recherche millénaire qui évolue encore actuellement. La nécessité de se dégager du langage naturel est apparue dès l'Antiquité avec la conscience des difficultés qu'il représentait. Les siècles qui ont suivi ont permis de longues réflexions sur la structure idéale des classements hiérarchiques, des arbres, afin de définir au mieux chaque Chose du monde. Cette réflexion constante a conduit à la constatation des limites de cette représentation hiérarchique du monde. Pourtant, Boèce l'avait déjà évoqué avec la création infinie d'arbres selon le contexte.

L'influence de l'arbre dans les systèmes documentaires et dans les référentiels est importante puisqu'elle est à l'origine de l'ensemble des vocabulaires contrôlés comme les *thesauri*. Comme les arbres, les vocabulaires contrôlés portent difficilement du sens au-delà des termes qui le composent : la hiérarchie ne rentre pas dans la définition sémantique des termes.

La possibilité infinie d'arbres conduit au vertige des labyrinthes qui se réinventent sans cesse selon le lecteur et sa position dans le labyrinthe. Au XX^{ème} siècle, les labyrinthes trouvent leur application dans l'informatique et dans le Web, créant ainsi des référentiels à leur image, désordonné et en constante évolution selon les requêtes qui y sont effectuées. Ainsi, les liens qui unissent les termes des référentiels portent du sens et permettent une définition d'un terme accompagnée de sens obtenu par ces liens. Ce « modèle-réseau », comme le nomme Umberto ECO, atteint son apogée avec la création de Wikidata. Cette plateforme est à la fois une encyclopédie basée sur des entités, des propriétés et des valeurs, conformément aux principes du Web de données, mais elle est également un lieu centralisé depuis lequel il est possible de se rediriger vers d'autres ressources par un simple lien. Contrairement aux volontés universalistes de Thomas D'AQUIN ou des Lumières, il n'existe pas un référentiel unique qui puisse décrire et définir l'ensemble du monde, mais bien

autant de référentiels et de jeux de données que nécessaire, suivant les spécificités de chacun et leur domaine d'activité. Le graphe de données présent sur le Web a permis de relier ces multiples référentiels et de s'affranchir de la création d'un jeu de données unique représentant le monde : l'ensemble des données reliées du monde constitue un référentiel qui décrit seul le monde. Alors, une institution patrimoniale qui souhaite apporter une description précise à ses documents peut utiliser ce graphe de données et ne conserver qu'un point d'entrée avec son identifiant.

À plus petite échelle, dans une institution ou une entreprise, la déconstruction des jeux de données et des référentiels en données reliées offre les mêmes avantages de navigation et de centralisation des connaissances du domaine d'activité.

En plus de cette évolution de la structure des référentiels, ce mémoire a montré comment les usages et les besoins ont évolué avec les référentiels. En effet, l'usage qui est fait d'un référentiel dirige sa structure jusqu'à la fin des années 2000. Ainsi, à l'INA, nous avons montré pour un même type de référentiel, celui des personnes physiques, mais pour deux métiers différents, la DJ et la DDCOL, les différences qui existent et qui trouvent leur origine dans les usages qui en sont faits : les niveaux de précision des informations et les structures divergent pour la description d'un même objet, une personne.

La nécessité de centraliser les données des diverses bases de données, et de leur redonner leur cohérence, a permis le lancement du *Lac de données*. La BnF ou bien le Centre Pompidou virtuel ont également mené à bien cette transformation du système documentaire dans la dernière décennie, afin de faciliter les usages faits des données. En effet, le processus de création du référentiel est retourné : le modèle de données est pensé selon la donnée qui est stocké, au lieu de partir des usages. Ce renversement permet de s'abstraire des usages qui seront faits des données pour proposer à l'avenir autant de services adaptés aux besoins des utilisateurs, à partir des mêmes données.

Ce changement dans la manière de concevoir un modèle de données provoque un déplacement du référentiel au sein des systèmes documentaires. Les *thesauri* et autres vocabulaires contrôlés étaient considérés comme de simples aides à la description : peu de termes étaient associés aux notices ce qui rendait la description sommaire et peu efficace pour la valorisation des documents. Le référentiel n'est qu'un outil de contrôle, comme peuvent l'être les règles de catalogage.

L'apparition du Web de données permet un enrichissement des données des institutions à partir d'un identifiant. Cependant, si cet enrichissement permet d'avoir plus d'informations concernant une entrée d'un thésaurus, il ne permet pas d'enrichir la description d'un document. L'utilisation du Web de données est alors subordonnée à la description préalable qui a été réalisée avec les référentiels internes.

L'impact du LED est important puisqu'il permet d'établir un modèle de données global

qui peut accueillir tout type de données sous la forme de nœuds et de liens. Avec ce LED, l'intelligence artificielle peut désormais décrire automatiquement chaque document avec précision, selon plusieurs outils de reconnaissance de textes, d'images ou de sons. Les descriptions créées sont alors plus complètes, et les liens vers les concepts du LED ou les entités du LOD se multiplient et accroissent la quantité d'informations qui sera disponible pour l'utilisateur final.

Les référentiels, bien qu'éclatés en données et en liens, seront toujours nécessaires aux institutions et aux entreprises pour la description contrôlée de leurs documents. L'évolution de leur forme, de leur structure et de leurs usages s'est produite sur plusieurs millénaires, et connaît de profonds bouleversements depuis les années 2000. Cette évolution est constante et nous pouvons imaginer que les futurs référentiels seront des référentiels partagés sur le Web certes, mais encore plus centralisés. Cette centralisation doit avoir lieu au-delà des limites longtemps perçues entre les institutions patrimoniales et le reste du Web, afin de décloisonner les référentiels et de les intégrer pleinement dans le Web sémantique.

Annexes

Annexe A

Les index de la Renaissance, termes contrôlés et classification alphabétique (les index de l'*Alsatia Illustrata* de Jean-Daniel Schoepflin)

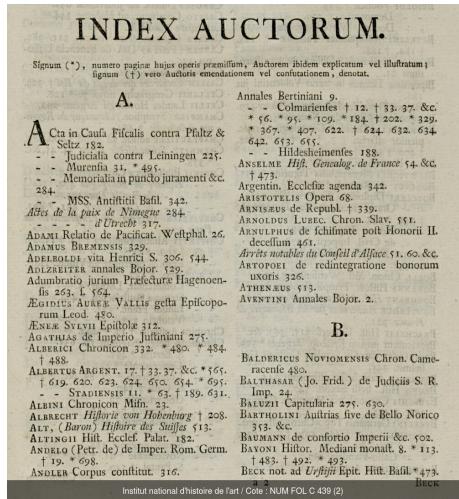


FIGURE A.1 – Index auctorum

Extraits des deux index de l'œuvre de Jean-Daniel SCHOEPFLIN [Source : <http://bibliotheque-numerique.inha.fr/idurl/1/12532>, p.804 et 813]

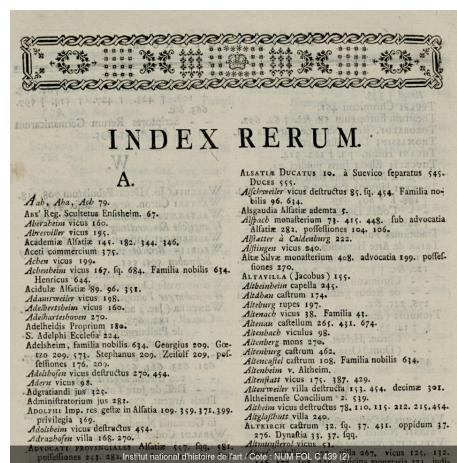


FIGURE A.2 – Index rerum

Annexe B

Les différents types d'interopérabilité

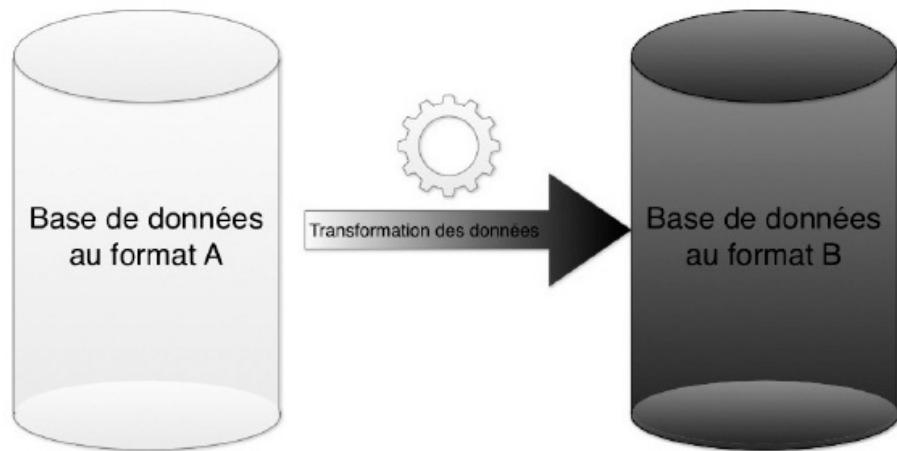


FIGURE B.1 – L'interopérabilité par conversion et copie [Source : bermes_2_2013]

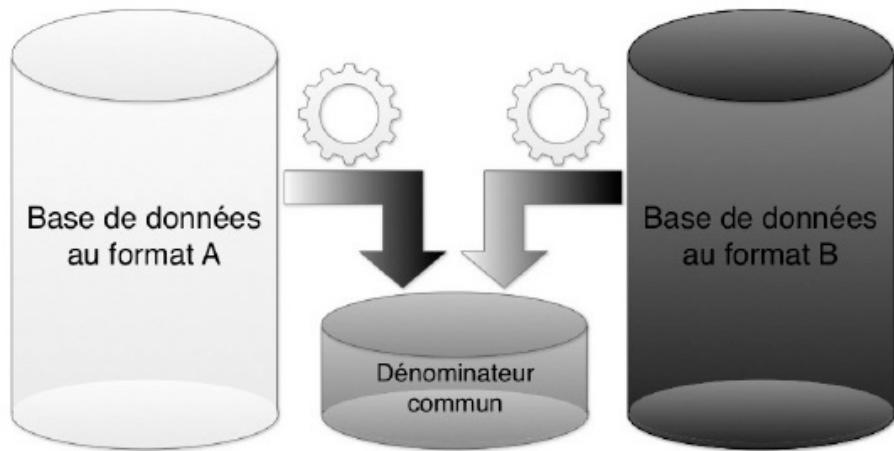


FIGURE B.2 – L’interopérabilité par le plus petit dénominateur commun [Source : bermes_2_2013]

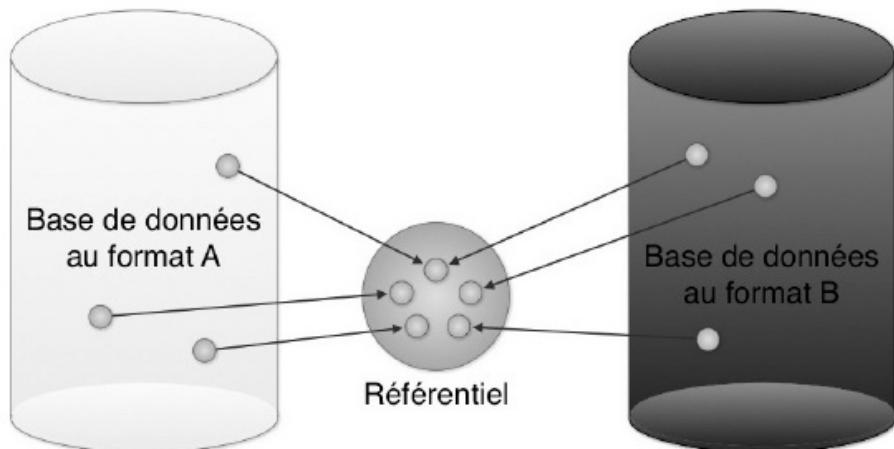


FIGURE B.3 – L’interopérabilité de la roue et de l’essieu [Source : bermes_2_2013]

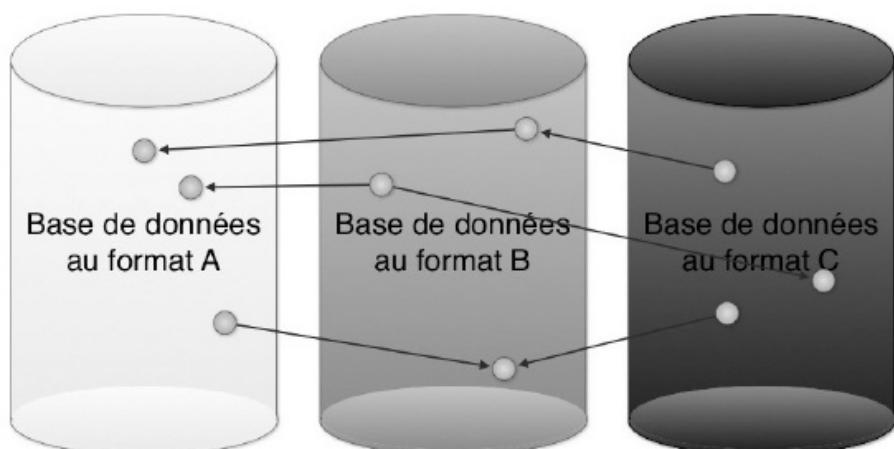


FIGURE B.4 – L’interopérabilité par parcours de liens [Source : bermes_2_2013]

Annexe C

Les types de données présents dans les bases de données de l'INA et leur rôle

Du texte libre



- **Texte court** : identifiant, titre propre, titre de la collection...
- **Texte long** : résumé, dispositif, notes...
- **Chiffres** : données d'audience, numéro d'émission, de saison...

Des données contrôlées



- **Description du contenu lui-même** : genre, thématique, descripteurs, génériques
- **Description des particularités du contenu** : langue, couleur, origine des images...
- **Description des particularités des événements** : nature de production, chaîne, nom du producteur

FIGURE C.1 – Les types de données présents dans les bases de données de l'INA [Source : ROCHE-DIORÉ (Axel), *Atelier transmission des connaissances*, 20 janv. 2020, p.6]

Annexe D

Les bases de données de la DDCOL de l'INA

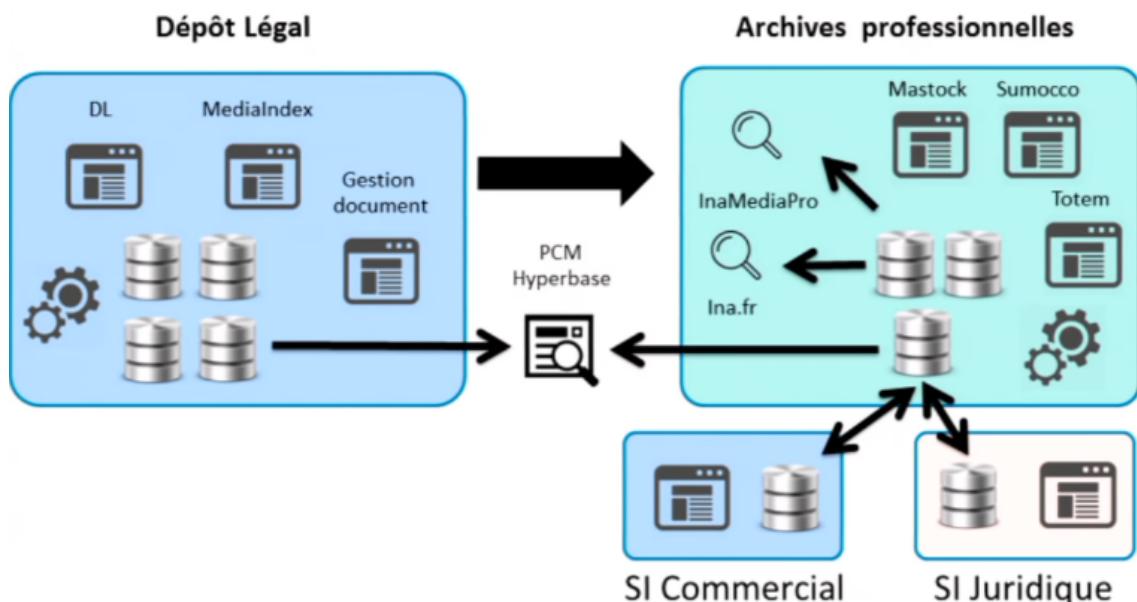


FIGURE D.1 – Les bases de données de la DDCOL de l'INA [Source : POUPEAU (Gautier), *Rassembler les métadonnées des collections de l'INA*, 11 févr. 2019, URL : <https://www.youtube.com/watch?v=KY0zoRPks8Q> (visité le 07/09/2020)]

Annexe E

Le thésaurus de noms communs de l'INA

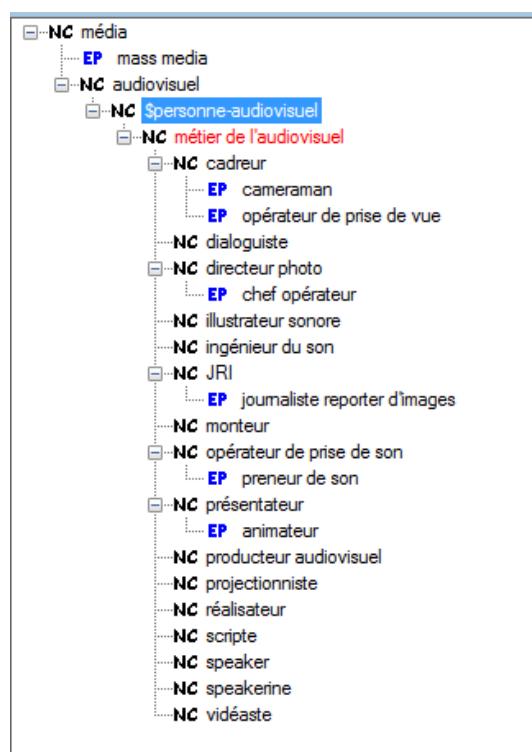


FIGURE E.1 – Extrait du thésaurus de noms communs de l'INA autour du terme « Cadreur »

Annexe F

Aligner les fonctions de « Journaliste » des notes qualité avec le thésaurus des noms communs de l'INA

Fonction normalisée des notes qualité	Équivalent du thésaurus
journaliste	journaliste
journaliste audiovisuel	journaliste
journaliste tv	journaliste
journaliste	journaliste
journaliste audiovisuel	
journaliste specialiste sant	
jounraliste	
jouornaliste	
jouranliste	
journalise	
journalisme	
journalitse	
journalliste	
journalsite	
jourrnaliste audiovisuel	

FIGURE F.1 – Résultat de l'alignement des journalistes avec le thésaurus des noms communs

Annexe G

Les captations directes réalisées par l'INA au titre du dépôt légal

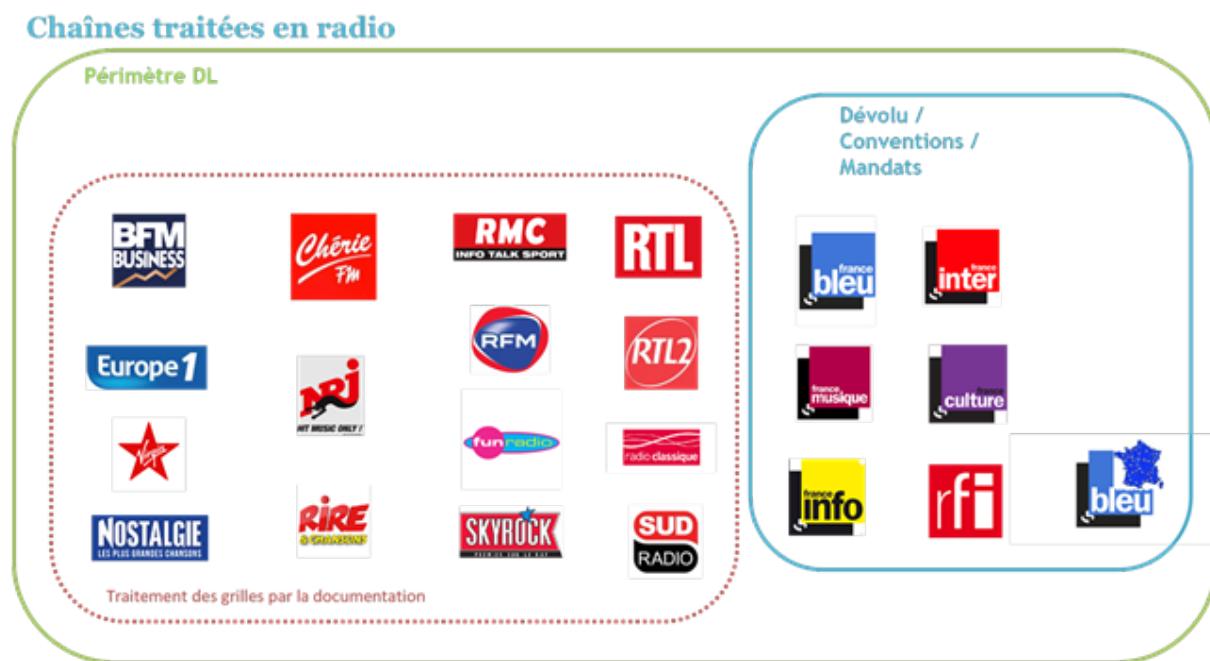


FIGURE G.1 – Stations de radio captées au titre du dépôt légal [Source : Communication électronique de l'entreprise « La collecte et le catalogage » du 26 mai 2020]

Chaînes traitées en télévision

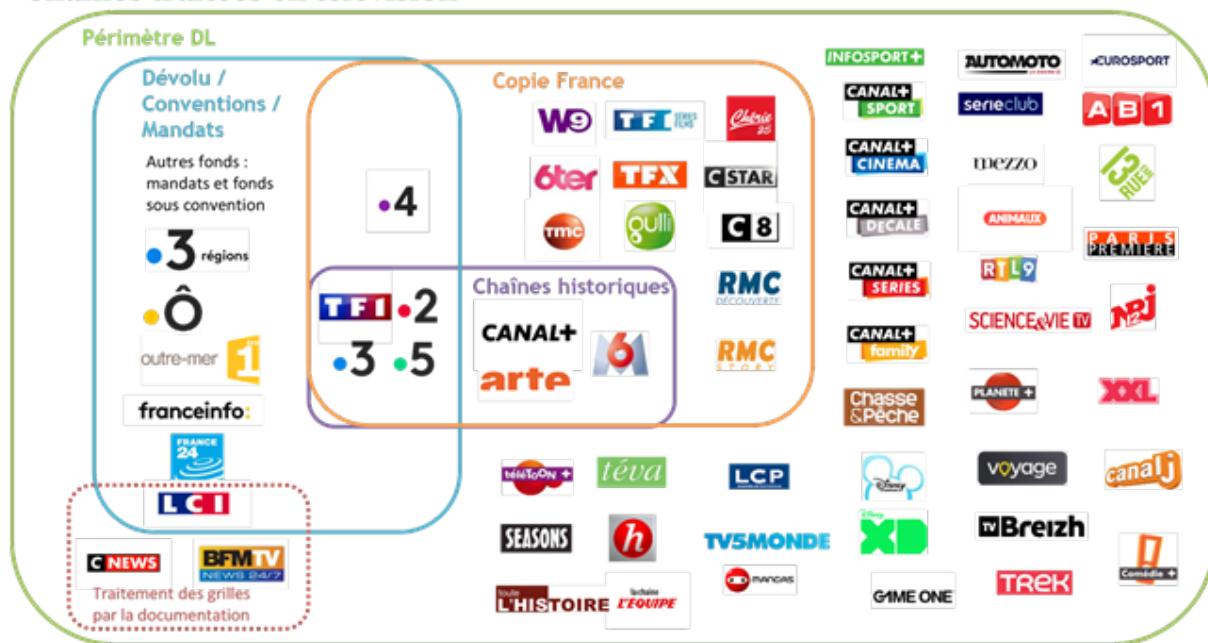


FIGURE G.2 – Chaînes de télévision captées au titre du dépôt légal [Source : Communication électronique de l’entreprise « La collecte et le catalogage » du 26 mai 2020]

Annexe H

Les fournisseurs externes de données de l'INA

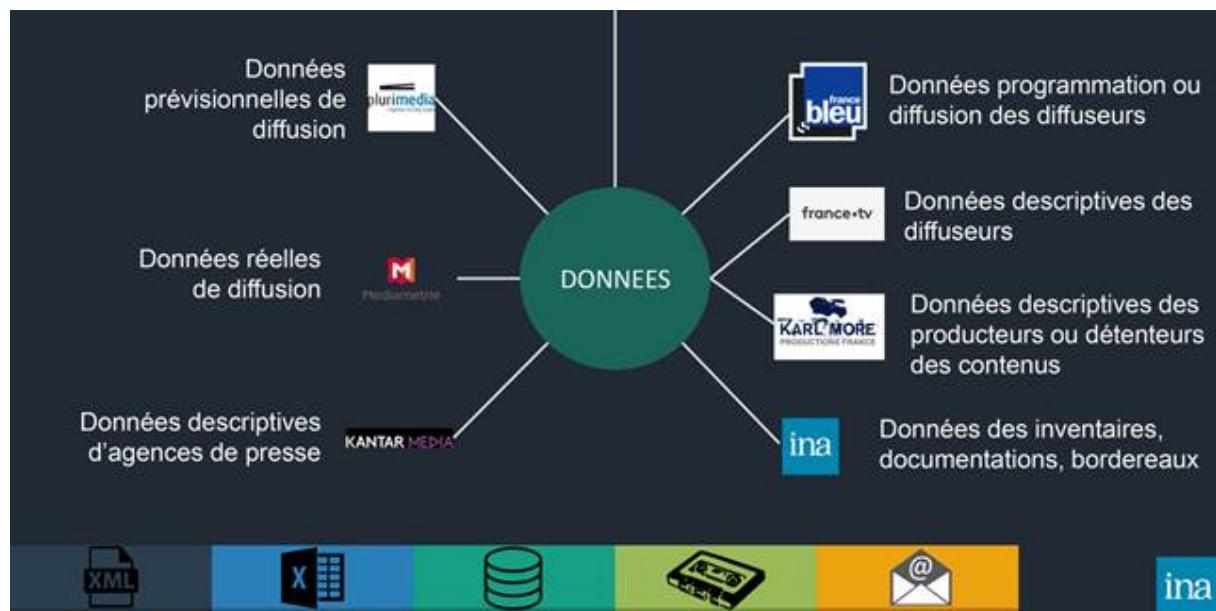


FIGURE H.1 – Les fournisseurs extérieurs de données de l'INA [Source : Communication électronique de l'entreprise « La collecte et le catalogage » du 26 mai 2020]

ID INA	CHAINE CODE	DOC TITRE COLLECTION	DOC TITRE PROPRE
550274 010 3	N12	The Big Bang Theory	La phobie de Sheldon
DOC DATE	DOC HEURE DEBUT	DOC HEURE FIN	DOC NUMERO EPISODE
20160312	124404	130424	S5-9
DOC NOMBRE EPISODES	DOC DUREE	DOC REPERAGE	DOC COULEUR
24	00202000	12440400	C
DOC TAUX MA	DOC PART DA	DOC TAUX MOYEN 15	DOC PART MOYEN 15
0.7	2.9	0.7	3
DOC TAUX MOYEN FEMME	DOC TAUX MOYEN HOMME	DOC DATE DERN MODIF	DOC LIEN REDIFFUSION
0.6	0.9	20171016	S :345421.025
DOC DATE CREATION	MEDIAMETRIE REF	ANNEE PRODUCTION	IMEDIA REF
20160315	56881	2011	35858990

TABLE H.1 – L'apport des données de Médiamétrie dans la description effectuée au DL
[Source : extrait de la base de données DLSAT de l'INA]

Annexe I

La constellation du Linked Open Data

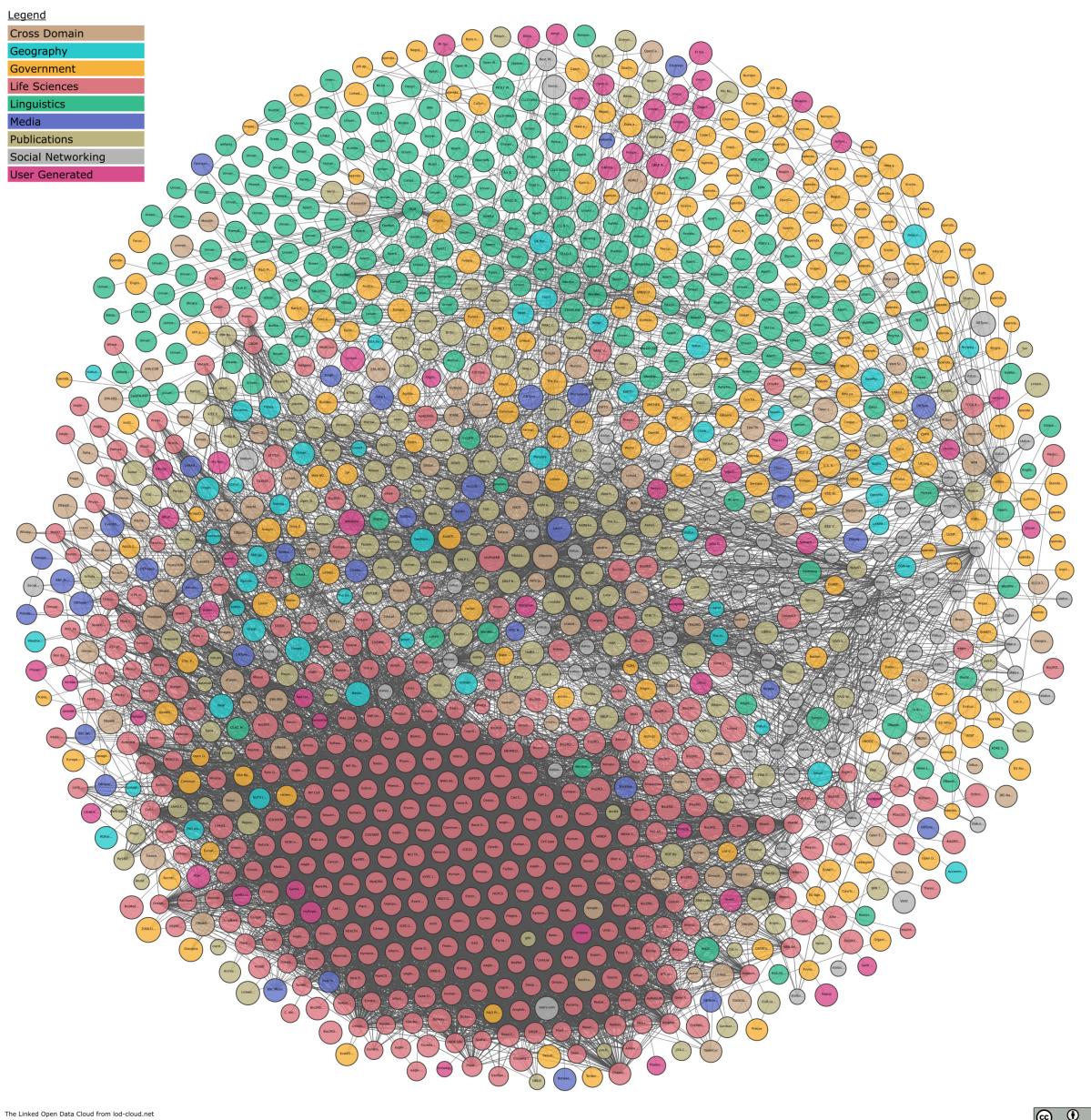


FIGURE I.1 – La constellation du Linked Open Data en juillet 2020 [Source : *The Linked Open Data Cloud*, The Linked Open Data Cloud, 27 juil. 2020, URL : <https://www.lod-cloud.net/> (visité le 27/07/2020)]



Annexe J

Repenser la place du référentiel

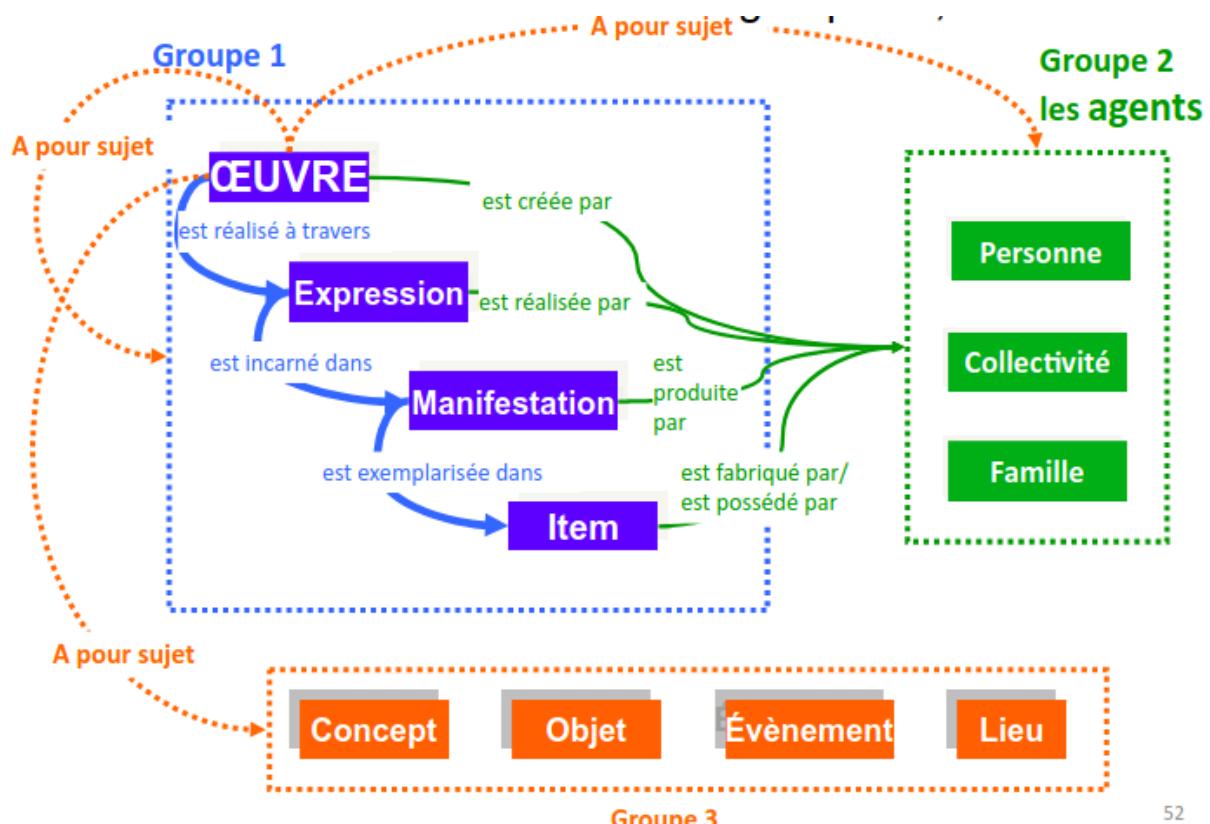


FIGURE J.1 – Le modèle FRBR [Source : BÉNÉZET (Joly), *Participer au Web de données*, 2015, URL : <https://slideplayer.fr/slide/3213771/> (visité le 09/09/2020), s.52]

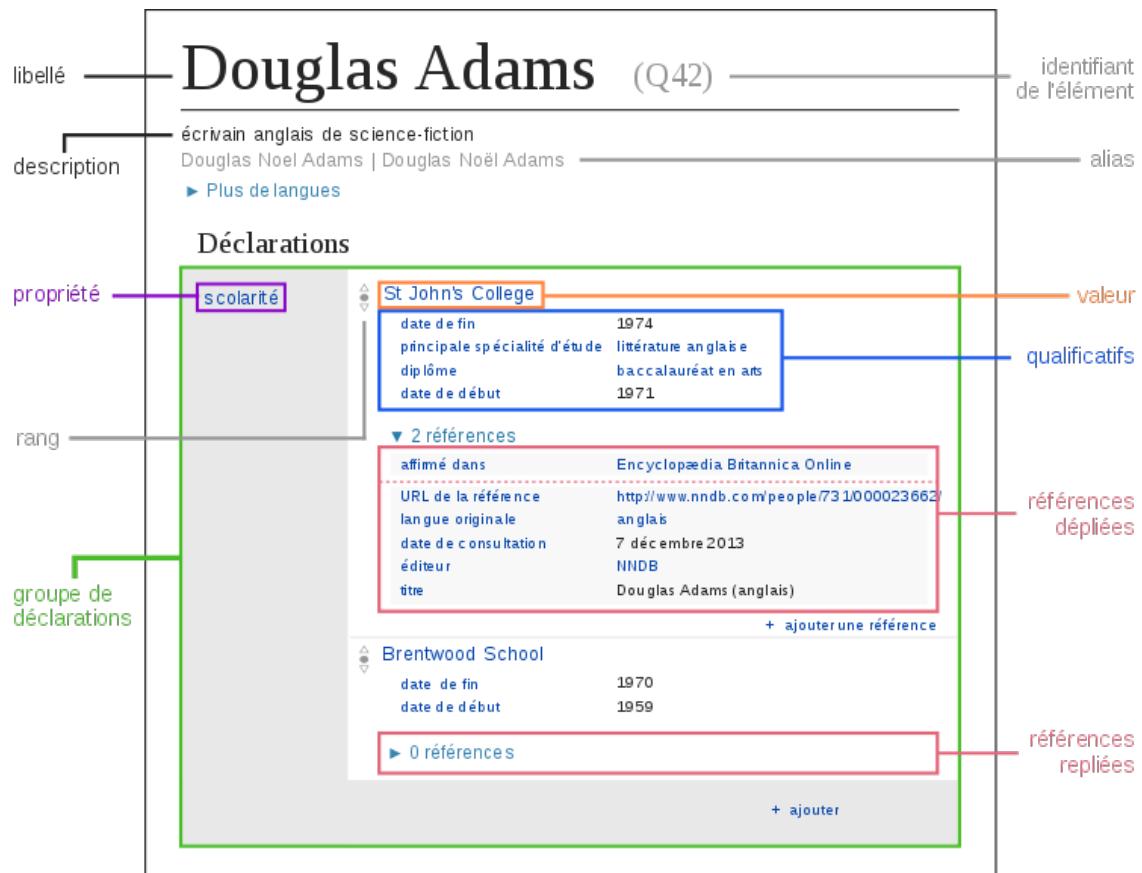
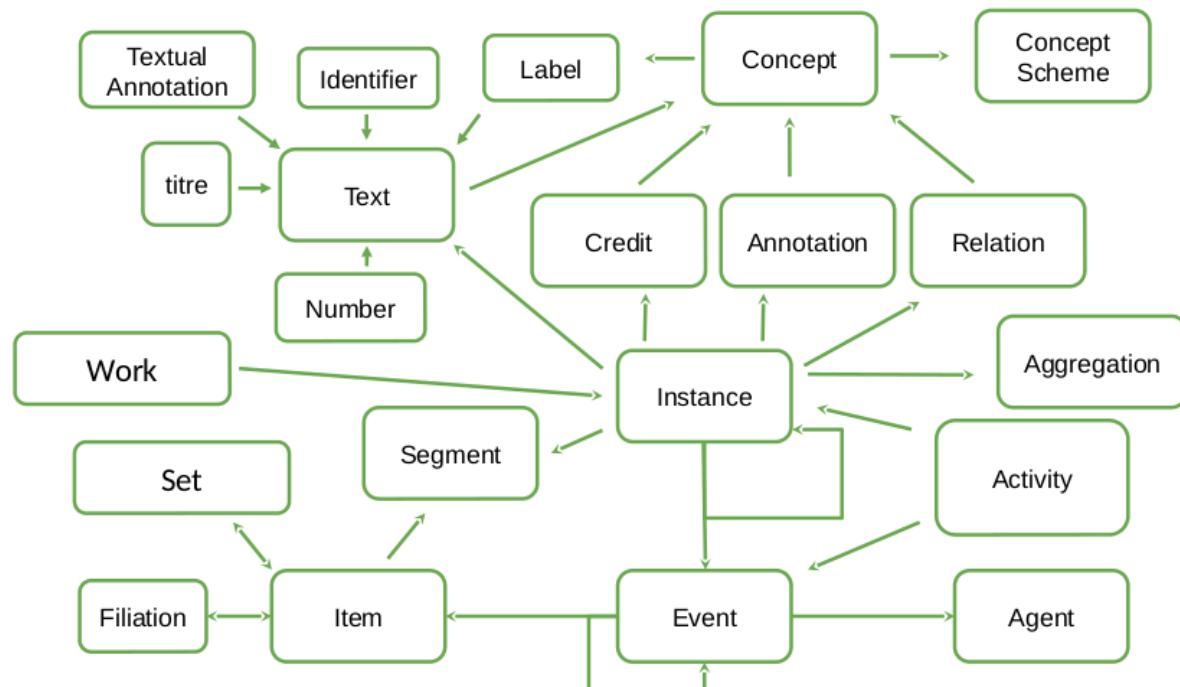


FIGURE J.2 – Le modèle de données de Wikidata [Source : Wikimédia]

FIGURE J.3 – Le modèle de données du *Lac de données* à l'INA [Source : ROCHE-DIORÉ (Axel), Atelier transmission des connaissances, 20 janv. 2020, d.13]

Annexe K

Repenser l'infrastructure

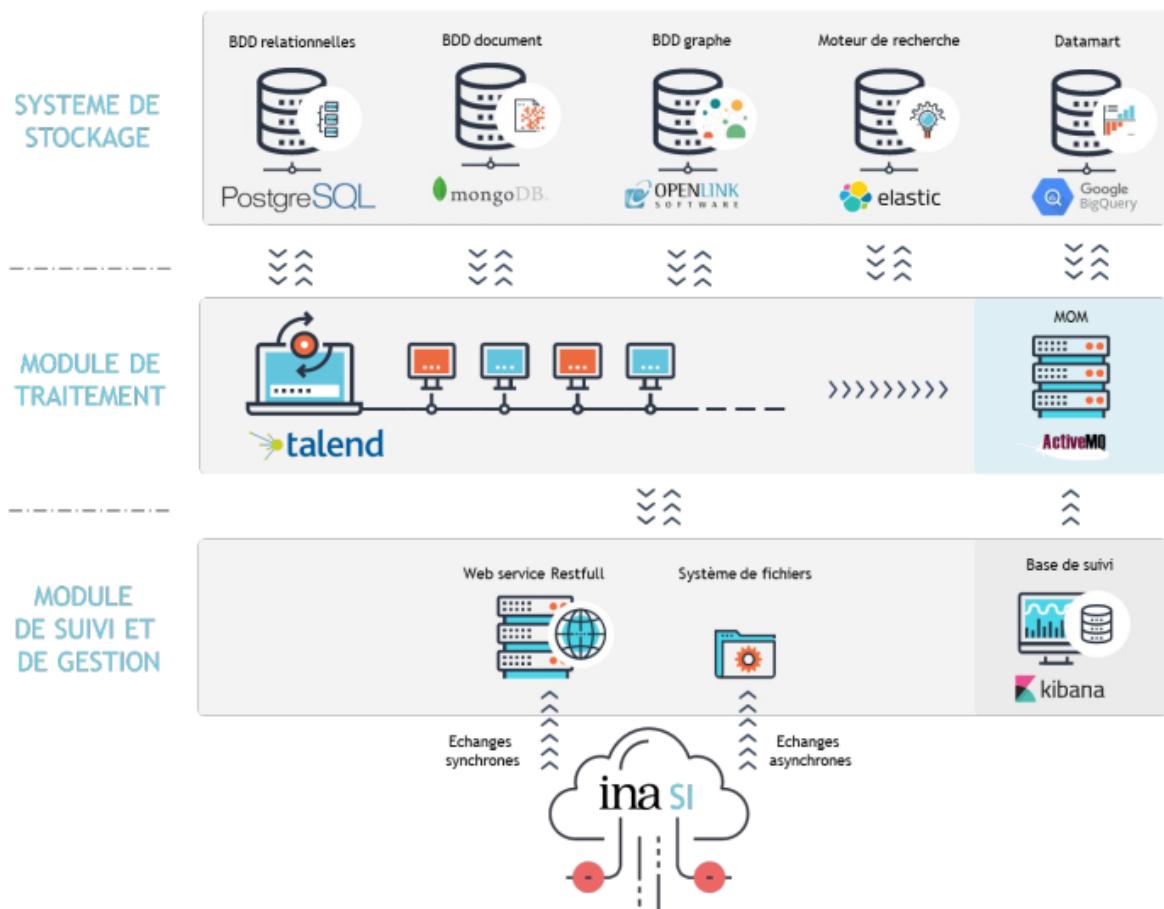


FIGURE K.1 – Schéma du *Lac de données* depuis le stockage jusqu'à l'accès aux données [Source : POUPEAU (Gautier), *Rassembler les métadonnées des collections de l'INA*, 11 févr. 2019, URL : <https://www.youtube.com/watch?v=KY0zoRPks8Q> (visité le 07/09/2020)]

Annexe L

Ontologies de haut et de bas niveaux

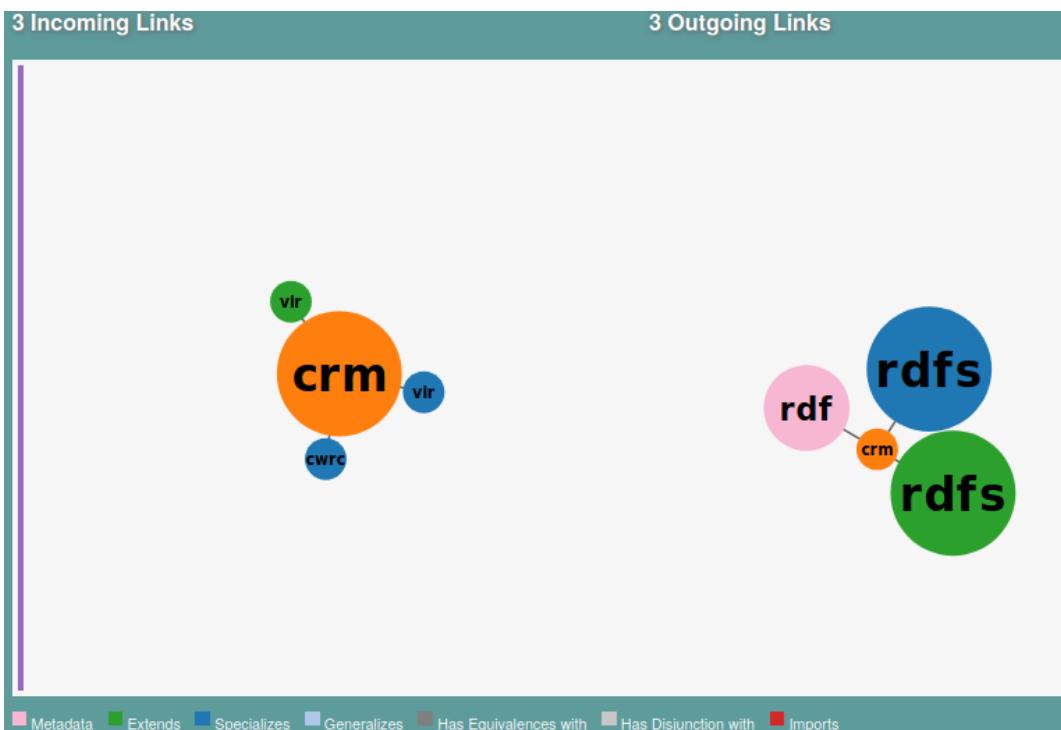


FIGURE L.1 – Une ontologie de haut niveau : représentation des ontologies utilisant celle du CIDOC-CRM (à gauche) et de celles utilisées par l'ontologie du CIDOC-CRM (à droite)
[Source : <https://lov.linkeddata.es/dataset/lov/vocabs/crm>]

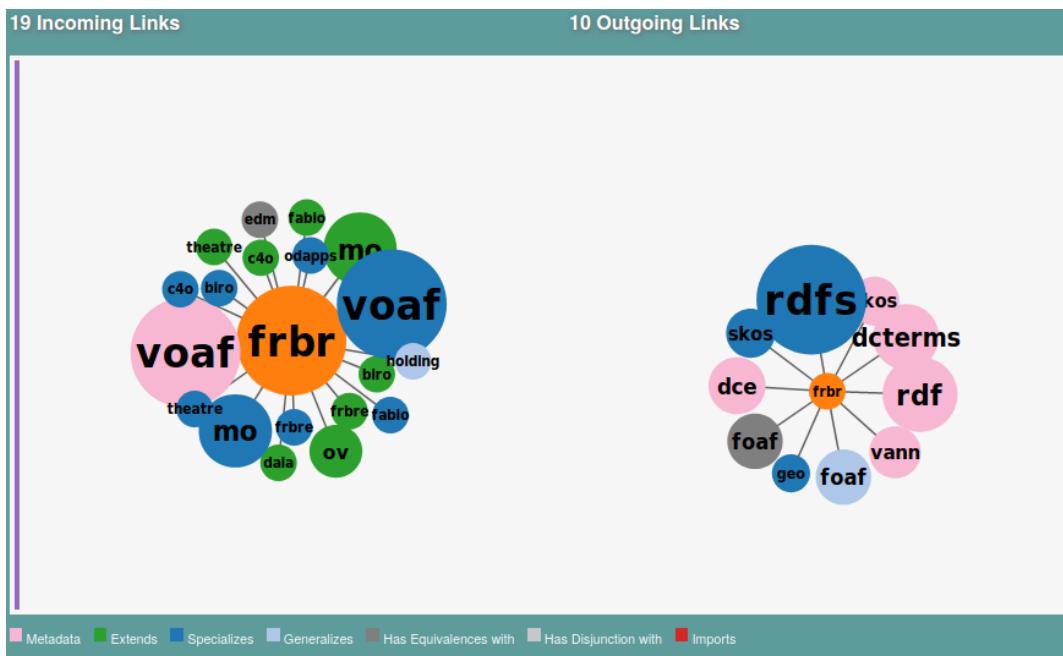


FIGURE L.2 – Une ontologie de domaine : représentation des ontologies utilisant celle du FRBR(à gauche) et de celles utilisées par l'ontologie du FRBR(à droite) [Source : <https://lov.linkeddata.es/dataset/lov/vocabs/frbr>]

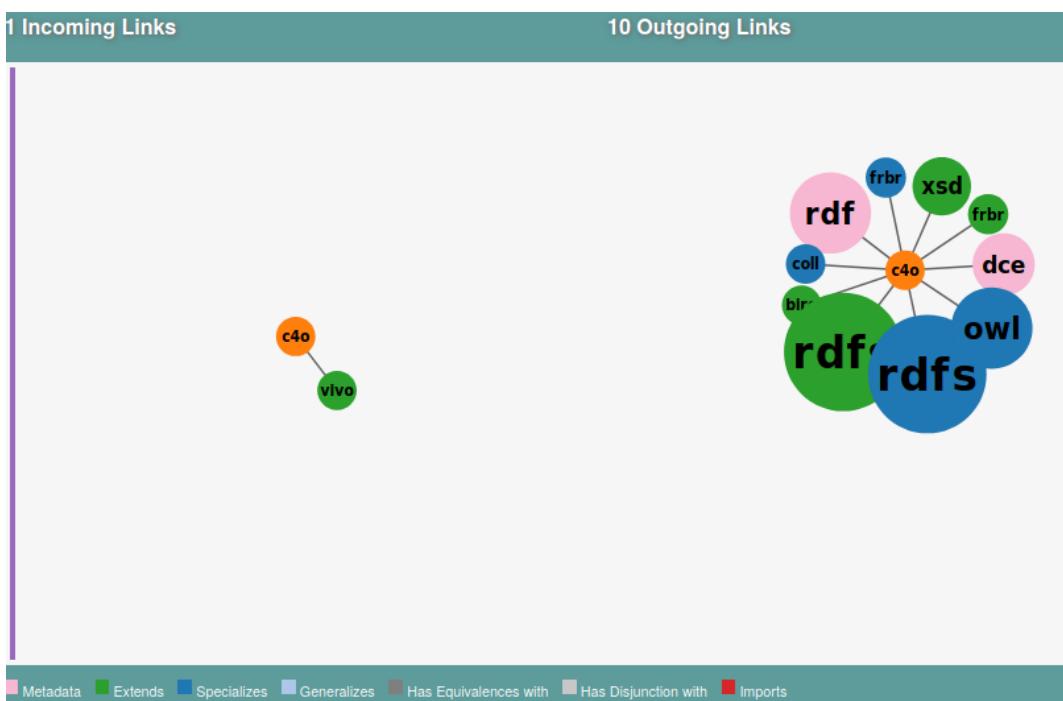


FIGURE L.3 – Une ontologie applicative : représentation des ontologies utilisant C4O (à gauche) et de celles utilisées par l'ontologie C4O (à droite) [Source : <https://lov.linkeddata.es/dataset/lov/vocabs/c4o>]

Annexe M

Labyrinthes



FIGURE M.1 – Le labyrinthe de Knos- *signe et l'interprétation*, trad. par Hélène sos [Source : ECO (Umberto), *De l'arbre Sauvage*, 1 t., Paris, 2010, ch. 1.5]
au labyrinthe : études historiques sur le signe et l'interprétation, trad. par Hélène Sauvage, 1 t., Paris, 2010, ch. 1.5]



FIGURE M.2 – Le labyrinthe d'Irrweg [Source : ECO (Umberto), *De l'arbre au labyrinthe : études historiques sur le signe et l'interprétation*, trad. par Hélène

au labyrinthe : études historiques sur le signe et l'interprétation, trad. par Hélène

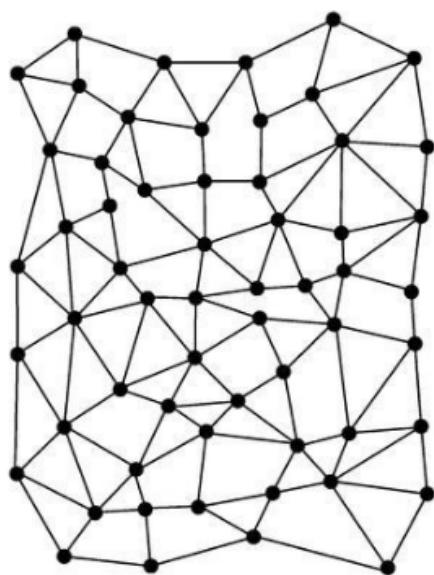


FIGURE M.3 – Le labyrinthe réseau [Source : Eco (Umberto), *De l'arbre au labyrinthe : études historiques sur le signe et l'interprétation*], trad. par Hélène Sauvage, 1 t., Paris, 2010, ch. 1.5]

Index des noms de référentiels

Dewey, 19

FOAF, 20

GEMET, 19

LCSH, 7

RAMEAU, 7

Table des figures

1.1	Anneau de synonymie du terme « Prisonniers » de RAMEAU	10
1.2	Anneau de synonymie du terme « Prisoners » de LCSH	10
2.1	Arbre porphyrien de l'homme avec les seuls attributs analytiques	16
2.2	Arbre porphyrien prenant en compte les différences	17
2.3	Infinitude de l'arbre de Porphyre	17
2.4	Extrait de l'arborescence de l'index de Pline L'ANCIEN	18
2.5	Classification des vocabulaires selon leur complexité	20
2.6	Le modèle taxonomique	21
2.7	Le modèle du thésaurus simple	21
2.8	Le modèle polyhiérarchique	22
2.9	Application du modèle polyhiérarchique	22
2.10	Relation d'équivalence	23
2.11	Relation d'association	23
2.12	Relation de hiérarchie	24
2.13	Modélisation d'une vedette de thésaurus	24
2.14	Données d'exemple de notes qualité avec la fonction de Réalisateur	27
2.15	Données d'exemple de notes qualité avec la fonction de Réalisateur, après normalisation des fonctions	27
2.16	Données d'exemple de notes qualité sans fonctions	27
2.17	Utilisation des synonymes pour l'alignement du terme « Cuisinier »	29
2.18	Gestion de la polysémie dans l'alignement du terme « Cuisinier »	29
3.1	Utilisation principale des référentiels conçus comme fournisseurs de clés . .	38
4.1	Modélisation des liens vers des référentiels externes présents dans la vedette « television » de LCSH	51
5.1	Modélisation simplifiée de SKOS	64
5.2	L'ontologie DC Terms	67
5.3	Le modèle de données de la BnF	68

6.1	Requête SPARQL de récupération des pays	77
6.2	Modélisation des entités Wikidata dont <i>Humoriste</i> est issu	80
6.3	Réutilisation d'un référentiel entre deux institutions	85
6.4	Utilisation commune d'un jeu de données du Web de données	86
7.1	La fédération entreprise par IdRef	97
7.2	Requête SPARQL pour récupérer les sous-classes de l'entité <i>Série Télévisée</i>	100
7.3	Requête SPARQL pour récupérer les identifiants des instances de la classe <i>Saison</i>	101
8.1	Modélisation des cinq entités du <i>Lac de données</i>	110
8.2	Modélisation globale des relations entretenues par les concepts	112
8.3	Modélisation du concept Mylène FARMER dans le <i>Lac de données</i>	113
8.4	Modélisation du concept Mylène FARMER dans le <i>Lac de données</i> et le LOD	114
8.5	Extraction des entités nommées d'un programme de France 2 avec DigInPix	117
8.6	Métadonnées associées à un document sur ina.fr	120
9.1	L'alignement des personnes de la DJ et de la DDCOL pour une jointure dans un tMap de Talend.	129
9.2	L'orchestration de la première étape dans l'ETL Talend.	130
9.3	Répartition des indices de confiance après les alignements entre les rééfren- tiels de personnes de la DJ et de la DDCOL	132
A.1	Index auctorum	141
A.2	Index rerum	141
B.1	L'interopérabilité par conversion et copie	143
B.2	L'interopérabilité par le plus petit dénominateur commun	144
B.3	L'interopérabilité de le roue et de l'essieu	144
B.4	L'interopérabilité par parcours de liens	144
C.1	Les types de données présents dans les bases de données de l'INA	145
D.1	Les bases de données de la DDCOL de l'INA	147
E.1	Extrait du thésaurus de noms communs de l'INA	149
F.1	Résultat de l'alignement des journalistes avec le thésaurus des noms communs	151
G.1	Stations de radio captées au titre du dépôt légal	153
G.2	Chaînes de télévision captées au titre du dépôt légal	154
H.1	Les fournisseurs extérieurs de données de l'INA	155

I.1	La constellation du Linked Open Data en juillet 2020	158
J.1	Le modèle FRBR	159
J.2	Le modèle de données de Wikidata	160
J.3	Le modèle de données du <i>Lac de données</i> à l'INA	160
K.1	Schéma du <i>Lac de données</i> depuis le stockage jusqu'à l'accès aux données .	161
L.1	Une ontologie de haut niveau : CIDOC-CRM	163
L.2	Une ontologie de domaine : FRBR	164
L.3	Une ontologie applicative : C4O	164
M.1	Le labyrinthe de Knossos	165
M.2	Le labyrinthe d'Irrweg	165
M.3	Le labyrinthe réseau	166

Liste des tableaux

6.1	Paramètres principaux du module <i>wbsearchentities</i> de l'API Wikibase	72
6.2	Paramètres principaux du module <i>wbgetentities</i> de l'API Wikibase	73
6.3	Paramètres principaux du module <i>wbgetclaims</i> de l'API Wikibase	73
6.4	Informations disponibles pour Howard Roberts à l'INA	74
6.5	Comparaison des informations disponibles pour Howard Roberts à l'INA et sur Wikidata	75
6.6	Comparaison des informations disponibles pour Howard Roberts à l'INA et sur Wikidata après un premier traitement	76
9.1	Points de contact entre les référentiels de la DJ et de la DDCOL	127
9.2	Scores attribués à chaque point de comparaison	128
H.1	L'apport des données de Médiamétrie dans la description effectuée au DL .	156

Table des matières

Résumé	iii
Remerciements	v
Liste des abréviations	vii
Introduction	ix
I CONTRÔLER. A la recherche de clés (années 1960 – fin des années 1990)	1
1 Le référentiel comme clé	5
1.1 Du langage libre au langage contrôlé : vers l'indexation	5
1.2 Une clé entre les données : les vocabulaires contrôlés	7
1.2.1 Contrôle de la forme des vedettes	8
1.2.2 Contrôle de la polysémie et de l'homographie	8
1.2.3 Contrôle de la synonymie	9
1.3 Une clé entre les jeux de données : l'interopérabilité par les fichiers d'autorité et les portails	9
1.3.1 La naissance des autorités par rétroconversion	11
1.3.2 Partager des vocabulaires : à la recherche de la meilleure interopérabilité	12
2 L'arbre, un vocabulaire contrôlé hiérarchique	15
2.1 L'arbre de Porphyre : origines et influences	15
2.1.1 L'arbre de Porphyre	16
2.1.2 L'encyclopédisme (Antiquité - Moyen-Âge) : la recherche d'un arbre global mimant le monde réel	18
2.1.3 Influences : une diversité de référentiels hiérarchiques	19
2.2 Le <i>thesaurus</i> , vocabulaire contrôlé hiérarchique le plus fréquent	20
2.2.1 Types de structure	21

2.2.2	Relations entre les termes	22
2.2.3	Utiliser la précoordination pour les relations complexes	24
2.3	Passer du texte libre à un vocabulaire contrôlé : aligner des notes qualité et un thésaurus de noms communs	25
2.3.1	Contrôler du texte libre	25
2.3.2	Aligner les extractions en langage naturel avec un thésaurus de noms communs	28
2.3.3	Classer selon le thésaurus	29
3	Les référentiels à l'INA	31
3.1	De multiples fonds à décrire	32
3.1.1	Les archives professionnelles	32
3.1.2	Les fonds issus du dépôt légal	32
3.2	Un système documentaire pluriel répondant aux besoins	33
3.2.1	Les bases de données du dépôt légal (DL)	34
3.2.2	Les bases de données des archives professionnelles (DA)	35
3.2.3	La base de données juridique (DJ)	35
3.3	Multiplication des sources de données et des référentiels	36
II RELIER. Vers le partage de référentiels communs (début des années 2000 – milieu des années 2010)		39
4 Le web de données : une exposition commune des référentiels		43
4.1	Le web de données : naissance et principes	44
4.1.1	Créer un modèle de données nativement compatible avec le web : le web de données	44
4.1.2	Inventer un format d'échange compatible avec ce modèle de données : RDF	47
4.2	La mise en commun de référentiels au service des institutions	48
4.2.1	L'adoption du Web de données en institutions patrimoniales	48
4.2.2	Utiliser des vocabulaires de valeurs	49
4.2.3	Créer des passerelles entre les référentiels	50
4.3	Vers la fin de la notion de référentiels ?	52
4.3.1	Quand tout devient un potentiel référentiel	52
4.3.2	Vers une uniformisation internationale de la donnée sur le Web et l'adoption de RDF comme format de production	54
5 Partager des structurations similaires de jeux de données par les classes et les propriétés : les ontologies, grammaires communes mais spécifiques		57

5.1	L'ontologie, un vocabulaire structurant	58
5.1.1	Origines de l'ontologie informatique	58
5.1.2	Des ontologies différentes	59
5.1.3	Les principes de l'ontologie	60
5.2	Des <i>Knowledge Organization Systems</i> (KOS) à SKOS : vers l'interopérabilité syntaxique	62
5.2.1	Distinguer les systèmes d'organisation de la connaissance des ontologies	62
5.2.2	SKOS : exposer les systèmes d'organisation de la connaissance sur le Web de données	63
5.3	Les ontologies dans le Web sémantique	65
5.3.1	Décrire des ontologies en RDF : RDFS et OWL	65
5.3.2	Utilisation des ontologies en institutions	66
6	Relier ses données à Wikidata : l'exemple de l'alignement des personnes physiques de l'INA	69
6.1	Effectuer des requêtes sur Wikidata	70
6.1.1	La structure des données de Wikidata	70
6.1.2	Le SPARQL-EndPoint	71
6.1.3	L'API Wikibase	72
6.2	Aligner des personnes depuis des données contrôlées	73
6.2.1	Choix des déclarations des entités de Wikidata	74
6.2.2	Adapter les données contrôlées pour les valeurs des déclarations	75
6.2.3	Effectuer l'alignement par de multiples requêtes	76
6.3	Aligner des personnes depuis du texte libre	78
6.3.1	Aligner les fonctions avec Wikidata	78
6.3.2	Utiliser la hiérarchie de Wikidata pour aligner les personnes	79
6.4	Comprendre les limites	81
6.4.1	Les raisons de l'absence d'alignement	82
6.4.2	Les limites du Web sémantique	83
III	CENTRALISER. Le référentiel, clé de voûte et pivot (depuis le milieu des années 2010)	87
7	Les labyrinthes comme réseaux de données et de liens	91
7.1	Du modèle encyclopédique aux graphes de données	91
7.1.1	Vers les labyrinthes (Renaissance)	92
7.1.2	Des labyrinthes aux graphes	93
7.2	Des labyrinthes de relations et d'identifiants : les hubs de liens	95

7.2.1	De la décentralisation des référentiels à leur recentralisation dans le Web de données	95
7.2.2	Apparition des hubs de liens et d'identifiants	96
7.2.3	Les hubs de liens et d'identifiants réceptacles de données	98
7.3	Wikidata comme hub de liens : aligner les fictions et les séries de l'INA avec Wikidata	99
7.3.1	Enrichir ses données avec des identifiants plutôt qu'avec des textes	99
7.3.2	Aligner des fictions et des séries avec Wikidata	100
7.3.3	Les difficultés posées par les langages naturels	102
8	Le <i>Lac de données</i> de l'INA : le référentiel au centre du modèle	105
8.1	Application des principes du Web de données aux systèmes documentaires : le Linked Enterprise Data	105
8.1.1	Permettre l'interopérabilité au sein des institutions	106
8.1.2	Repenser le système documentaire	107
8.1.3	Le positionnement du référentiel	107
8.2	Le <i>Lac de données</i> de l'INA : un repositionnement du référentiel au centre du modèle de données	109
8.2.1	Le <i>Lac de données</i> , un modèle basé sur des classes d'entités	109
8.2.2	La place des concepts	111
8.2.3	Le <i>Lac de données</i> comme un LED : une infrastructure unique	112
8.3	Perspectives d'utilisation	115
8.3.1	Permettre l'intégration des données issues de la description et de la segmentation de vidéos dans le <i>Lac de données</i> : réutilisation des concepts et enrichissement des métadonnées	116
8.3.2	Faciliter et améliorer le catalogage des documents de l'INA par l'extraction automatique de données	118
8.3.3	Améliorer la valorisation des documents et offrir une meilleure expérience utilisateur	119
9	Centraliser les référentiels de l'INA dans le <i>Lac de données</i> : l'exemple de l'alignement de deux référentiels de personnes physiques	123
9.1	Des jeux de données différents en de multiples points	124
9.1.1	Enjeux	124
9.1.2	Points de contact	125
9.1.3	Divergences	125
9.2	Établir une méthodologie particulière d'alignement	127
9.2.1	Créer un indice de confiance pour chaque alignement	128
9.2.2	Des étapes exclusives	129
9.3	Des résultats à la hauteur des données initiales	131

9.3.1	Des résultats hétérogènes reflétant les multiples difficultés	131
9.3.2	Une supervision humaine nécessaire	132
Conclusion		135
Annexes		141
A	Les index de la Renaissance, termes contrôlés et classification alphabétique (les index de l' <i>Alsatia Illustrata</i> de Jean-Daniel Schoepflin)	141
B	Les différents types d'interopérabilité	143
C	Les types de données présents dans les bases de données de l'INA et leur rôle	145
D	Les bases de données de la DDCOL de l'INA	147
E	Le thésaurus de noms communs de l'INA	149
F	Aligner les fonctions de « Journaliste » des notes qualité avec le thésaurus des noms communs de l'INA	151
G	Les captations directes réalisées par l'INA au titre du dépôt légal	153
H	Les fournisseurs externes de données de l'INA	155
I	La constellation du Linked Open Data	157
J	Repenser la place du référentiel	159
K	Repenser l'infrastructure	161
L	Ontologies de haut et de bas niveaux	163
M	Labyrinthes	165
Index des noms de référentiels		167
Table des figures		169
Liste des tableaux		173
Table des matières		175