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

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

GEMET General Multilingual Environmental Thesaurus

INA Institut national de l'Audiovisuel

ISAN *International Standard Audiovisual Number*

LCSH Library of Congress Subject Headings

MARC MAchine-Readable Cataloging

OAI-PMH Open Archive Intiative Protocol for Metadata Harvesting

OCLC Online Computer Library Center

ORTF Office de la radio-télévision française

RAMEAU Répertoire d'autorité-matière encyclopédique et alphabétique unifié

UNIMARC UNIversal MAchine-Readable Cataloging

Introduction

> « Toutefois pour ne laisser cette quantité infinie ne la définissant point, [et] aussi pour ne jetter 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'intégrr 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)

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 langage [1] »

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 le 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$^{\text{ème}}$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 créé 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 alphabétique [2].

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$^{\text{ème}}$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 système [5] » 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 utilisateurs [6], 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 autorités [7] et des [8], 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 Celtica 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 Schoepflini, Colmariae [Colmar], 1751 (Collaction 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 synonymes [9]. 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'utilisation [11].

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.*

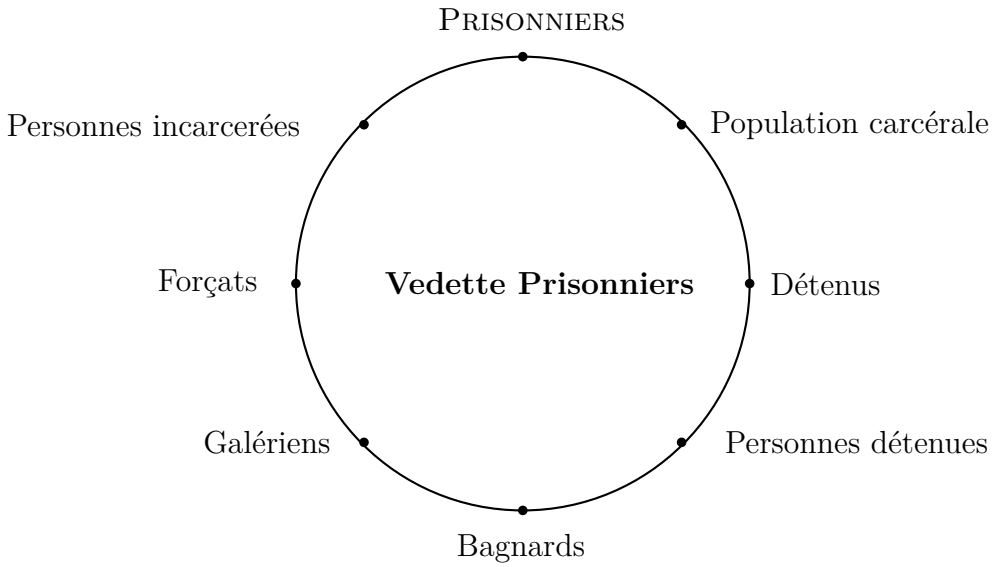


Figure 1.1 – Anneau de synonymie du terme « Prisonniers » de RAMEAU

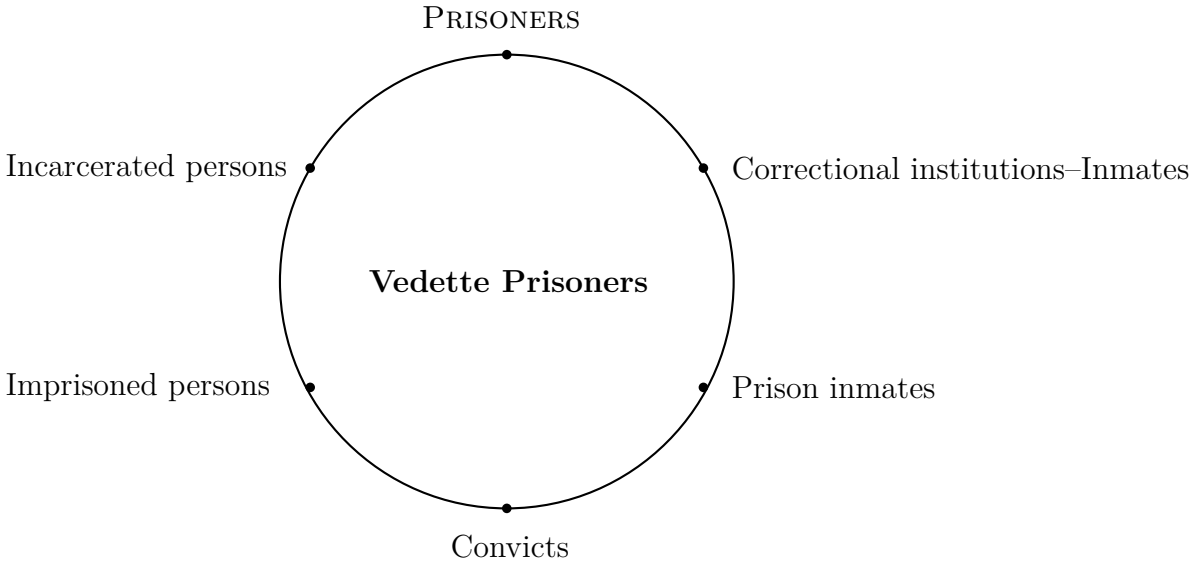


Figure 1.2 – Anneau de synonymie du terme « Prisoners » de LCSH

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 informatisé 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 créé de multiples renvois à l'intérieur du fichier physique ou informatique. L'arrivée des moteurs de recherche dans les années 2000 permet de supprimer

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. Emmanuelle Bermès, Antoine Isaac et Gautier Poupeau, "1. Du catalogue de bibliothèque aux données sur le Web : un changement de paradigme du côté de l'usager", *Bibliotheques* (, 2013), ISBN : 9782765414179 Publisher : Éditions du Cercle de la Librairie, p. 17-28, URL : `https://www-cairn-info.proxy.chartes.psl.eu/le-web-semantique-en-bibliotheque--9782765414179-page-17.htm` (visité le 31/07/2020).

ces différences de termes en indexant à la fois les formes retenues et les formes rejetées, permettant de trouver directement la vedette recherchée.

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 les 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 MARC [15].

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

14. Id., "2. Convergence et interopérabilité : vers le Web de données", *Bibliotheques* (, 2013), ISBN : 9782765414179 Publisher : Éditions du Cercle de la Librairie, 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).

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. E. Bermès, A. Isaac et G. Poupeau, "2. Convergence et interopérabilité…".

17. *Ibid.* Voir Annexe B : Les différents types d'interopérabilité

18. B. nationale de France, *Le protocole Z39.50…*

de manière à se connecter au portail correspondant (sui lui-même doit être connu de cet utilisateur) [19].

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 Intiative 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 système [20].

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 Core [21] 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 dénominteur 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.

19. Sylvie Dalbin, E. Bermès, A. Isaac, Romain Wenz, Y. Nicolas, Tayeb Merabti, Anila Angjeli, Thomas Francart, Lise Rozat, Pierre-Yves Vandenbussche, *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.chartes.psl.eu/revue-documentaliste-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.bnf.fr/fr/protocole-oai-pmh` (visité le 02/09/2020).

21. *Dublin Core Metadata Initiative*, URL : `https://dublincore.org/` (visité le 02/09/2020).

22. E. Bermès, A. Isaac et G. Poupeau, "2. Convergence et interopérabilité…". Voir Annexe B : Les différents types d'interopérabilité

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.

Chapitre 2

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 conserver [1]. »

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 depuis 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 service 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 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.psl.eu/l-ina--9782130545453.htm` (visité le 05/07/2020), p.45.

2.1 De multiples fonds à décrire

2.1.1 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 « Actualité 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.

2.1.2 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 NRJ [2].

2. En 2020, l'ensemble des stations captées au titre du dépôt légal par l'INA est décrit dans l'Annexe F : Les captations directes réalisées par l'INA au titre du dépôt légal (Figure F.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 captées [3].

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 2019 [4], 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.

2.2 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 documents [5], 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 propre 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 F : Les captations directes réalisées par l'INA au titre du dépôt légal (Figure F.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.chartes.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.

2.2.1 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 radio [6]. 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.

2.2.2 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 communs [7] et des personnes physiques et morales.

2.2.3 La base de donnés juridique (DJ)

La base de données « Adaje » de la DJ comprend l'ensemble des données permettant d'identifier et de rémunérer les ouvrants-droit [8] et les ayants-droit [9] 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 D : 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 DDCOL et la DJ présentées ici. Leur structure[11] 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.

2.3 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 forunies.

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 document, 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 multiples[12] et fournissent des données tant sur les programmes que sur les producteurs eux-mêmes :

— Les données prévisionnelles de diffusion de la télévision sont achetées auprès de la société Plurimédia[13]. Les fictions, les documentaires, les dessins animés, les

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

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

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

13. Voir `http://www.plurimedia.fr/`.

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 Médiamétrie [14], 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 Kantarmédia [15] ; pour les producteurs, les informations sont obtenues depuis la société Karl More Productions France.

—

14. Voir `https://www.mediametrie.fr/`.
15. Voir `https://www.kantarmedia.com/fr`.

Chapitre 3

L'arbre, un vocabulaire contrôlé hiérarchique

Nous l'avons évoqué (section 3.2 : 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.

3.1 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.

3.1.1 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 créé 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 genres [4]. Un arbre de Porphyre est par conséquent une succession de genres divisés en espèces qui deviennent elles-mêmes des genres.

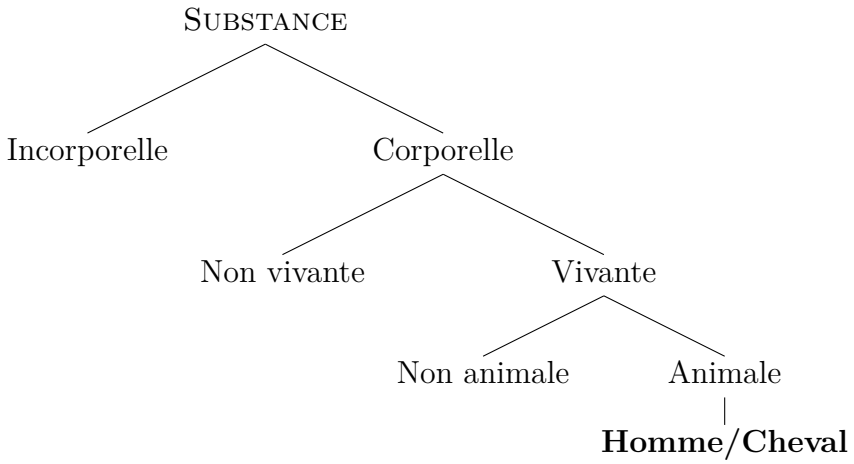


Figure 3.1 – Arbre porphyrien de l'homme avec les seuls attributs analytiques [d'après Eco (Umberto), *De l'arbre au labyrinthe : [études historiques sur le signe et l'interprétation]*, trad. par Hélène Sauvage, 1 t., Paris, 2010]

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 3.2 : 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 3.1 : Arbre porphyrien de l'homme avec les seuls attributs analytiques
5. *Ibid.*

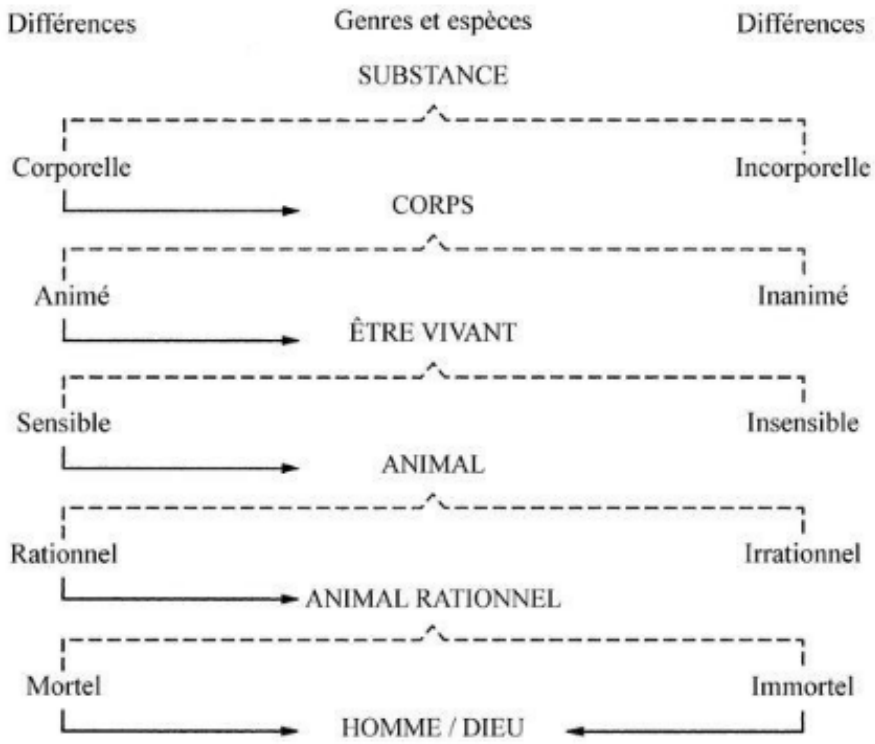


FIGURE 3.2 – Arbre porphyrien prenant en compte les différences [Source : ECO (Umberto), *De l'arbre au labyrinthe : [études historiques sur le signe et l'interprétation]*, trad. par Hélène Sauvage, 1 t., Paris, 2010]

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 3.3 : Infinitude de l'arbre de Porphyre).

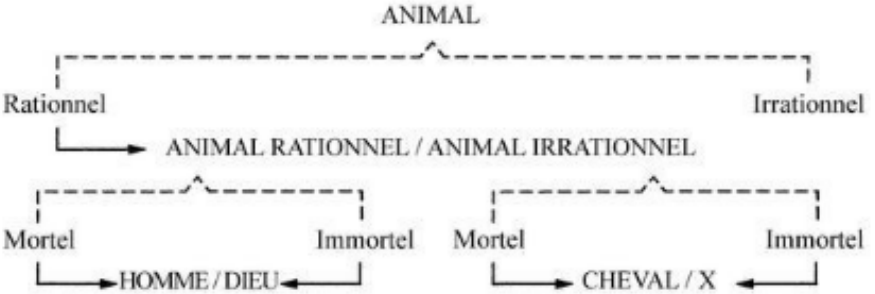


FIGURE 3.3 – Infinitude de l'arbre de Porphyre [Source : ECO (Umberto), *De l'arbre au labyrinthe : [études historiques sur le signe et l'interprétation]*, trad. par Hélène Sauvage, 1 t., Paris, 2010]

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 infini, 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.

3.1.2 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.

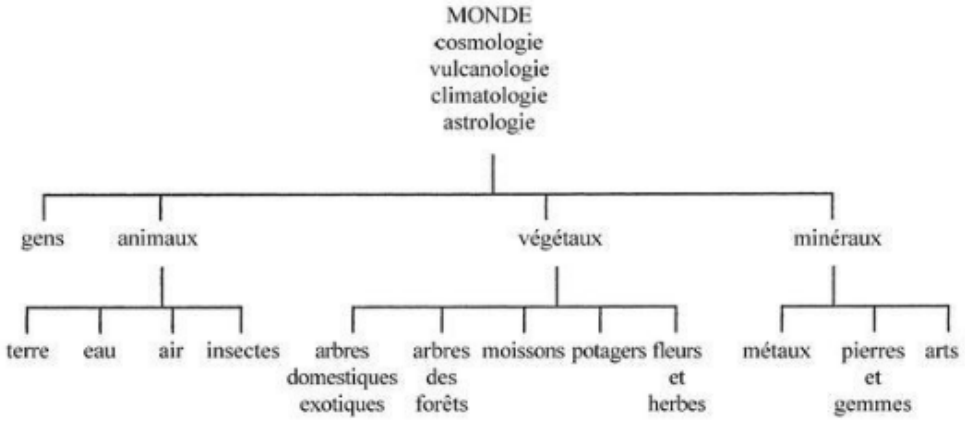


Figure 3.4 – Extrait de l'arborescence de l'index de Pline L'Ancien

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 monde [7]. À partir du XIII$^{\text{ème}}$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 3.4 : 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.

3.1.3 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 différences [9]. 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 nœud, 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 Rosenfeld et Peter Morville [10], 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 connaissance [12]. 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 3.2 : 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 Rosenfeld et Peter Morville [13] 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.

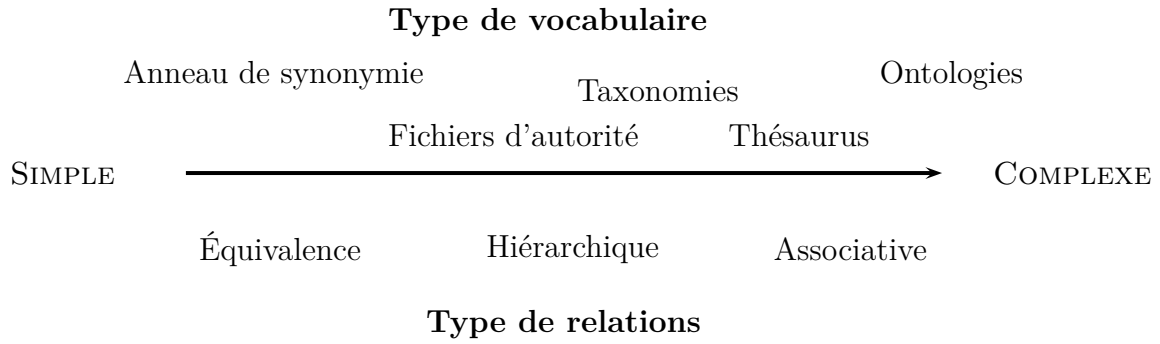


Figure 3.5 – Classification des vocabulaires selon leur complexité [d'après Rosenfeld (Louis), Morville (Peter) et Arango (Jorge), *Information architecture for the World Wide Web*, OCLC : 922954742, Beijing, 2015]

3.2 Le *thésaurus*, 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 3.5 : 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.

3.2.1 Types de structure

Le type de thésaurus le plus utilisé est celui constitué d'une hiérarchie simple [14]. L'INA possède un thésaurus de noms communs formé sur cette hiérarchie simple à unique ascendance [15], 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 taxinomie [16].

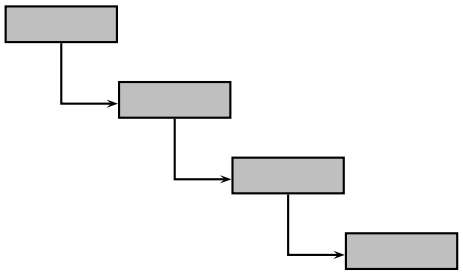


FIGURE 3.6 – Le modèle taxonomique

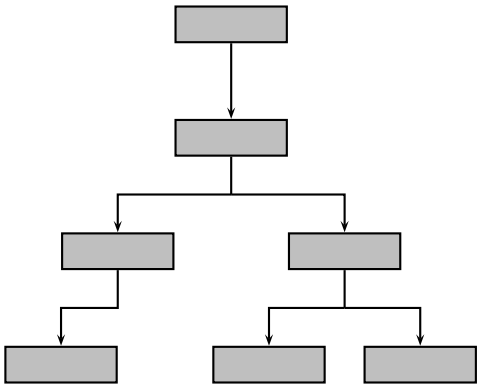


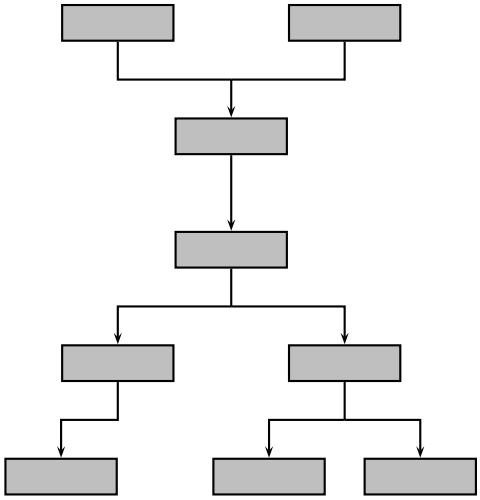
FIGURE 3.7 – Le modèle du thésaurus simple

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 précédemment [17], 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 D : Le thésaurus de noms communs de l'INA
16. Voir Figure 3.6 : Le modèle taxonomique et Figure 3.7 : Le modèle du thésaurus simple.
17. Voir section 3.1 : L'arbre de Porphyre : origines et influences.



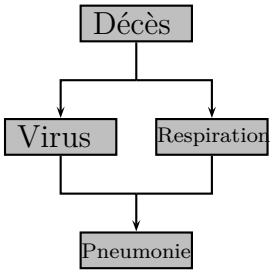


FIGURE 3.9 – Application du modèle polyhiérarchique

FIGURE 3.8 – 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.

3.2.2 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 relation [19] est une relation horizontale, d'égalité, comme dans l'anneau de synonymie. Dans l'Annexe D : 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 ».

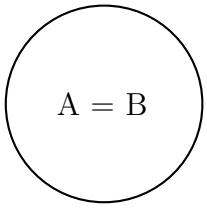


FIGURE 3.10 – Relation d'équivalence

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'association [20] 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 :

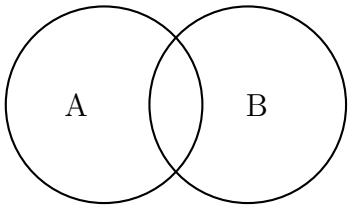


FIGURE 3.11 – Relation d'association

- 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 hiérarchique [21] 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.

19. Voir Figure 3.10 : Relation d'équivalence
20. Voir Figure 3.11 : Relation d'association.
21. Voir Figure 3.12 : Relation de hiérarchie.

3.2.3 Utiliser la précoordination pour les relations complexes

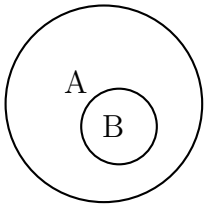


FIGURE 3.12 – Relation de hiérarchie

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 complètent [22]. 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.

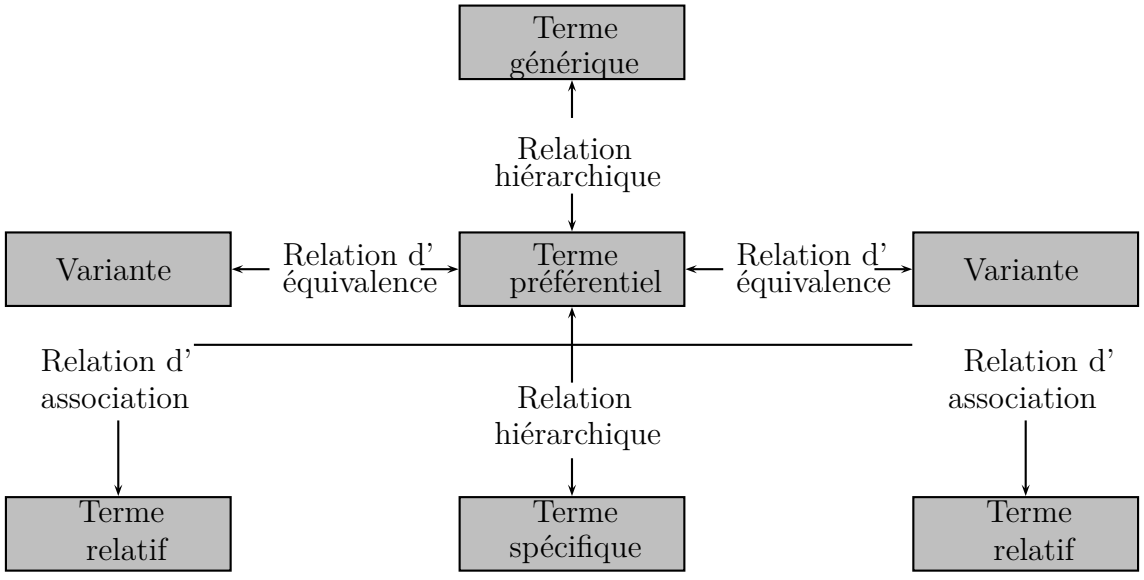


FIGURE 3.13 – Modélisation d'une vedette de thésaurus [d'après ROSENFELD (Louis), MORVILLE (Peter) et ARANGO (Jorge), *Information architecture for the World Wide Web*, OCLC : 922954742, Beijing, 2015]

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

3.3 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.

3.3.1 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 Loaad*) [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 3.14 montre une dernière normalisation à effectuer : la suppression des *stopwords*, effectuée en Figure 3.15.

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	scenariste
3	Desinateur, illustrateur, graphiste. France	desinateur
3	Desinateur, illustrateur, graphiste. France	illustrateur
3	Desinateur, illustrateur, graphiste. France	graphiste
4	Dessinatrice, animatrice. France	dessinateur
4	Dessinatrice, animatrice. France	animateur

FIGURE 3.14 – 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	scenariste
3	Desinateur, illustrateur, graphiste. France	desinateur
3	Desinateur, illustrateur, graphiste. France	illustrateur
3	Desinateur, illustrateur, graphiste. France	graphiste
4	Dessinatrice, animatrice. France	dessinateur
4	Dessinatrice, animatrice. France	animateur

FIGURE 3.15 – 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 fonctions[24]. 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 3.16 – Données d'exemple de notes qualité sans fonctions

24. Voir Figure 3.16 : Données d'exemple de notes qualité sans fonctions.

3.3.2 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 réalisé [25]. 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 synonymes [26], 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 jointures [27]. 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 exact [28].

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 réalisés [29].

Enfin, le cas de « Chef cuisinier » montre la nécessité d'utiliser le second terme de l'expression de la fonction [30] : 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 D : 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 E : Aligner les fonctions de « Journaliste » des notes qualité avec le thésaurus des noms communs de l'INA.

29. Voir Figure 3.17 : Utilisation des synonymes pour l'alignement du terme « Cuisinier »

30. Voir Figure 3.18 : 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		
2	Cuisinière. France	cuisinier	cuisinier	1
3	Cheffe cuisinière. France	chef cuisinier		
4	Cheffe cuisinier. Italie	chef cuisinier		
5	Chef de cuisine. France	chef cuisine	chef de cuisine	1

FIGURE 3.17 – 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 3.18 – Gestion de la polysémie dans l'alignement du terme « Cuisinier »

3.3.3 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 intermédiaires [31]. 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

Deuxième partie

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

Chapitre 4

Le web de données : une exposition commune des référentiels

Chapitre 5

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

Chapitre 6

Relier ses données à Wikidata

Troisième partie

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

Chapitre 7

Les labyrinthes comme réseaux de données et de liens

Chapitre 8

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

Chapitre 9

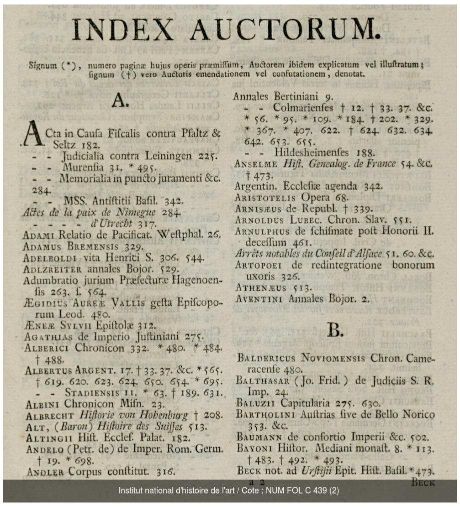
Centraliser les référentiels de l'INA dans le Lac de données : l'exemple de l'alignement de deux référentiels de personnes physiques

Conclusion

Annexes

Annexe A

Les index de la Renaissance, termes contrôlés et classification alphabétique (les index de l'*Alsatia Illustrata* de Jean-Daniel Schoepflin)



INDEX AUCTORUM.

Signum (*), numero paginæ hujus operis præmiſſum, Auctorem ibidem explicatum vel illuſtratum; ſignum (†) vero Auctoris emendationem vel confutationem, denotat.

A.

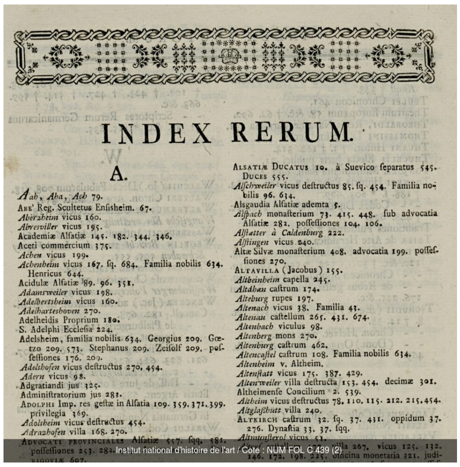
Acta in Cauſa Fiſcalis contra Pfaltz & Seltz 182.
- - Judicialia contra Leiningen 225.
- - Murenſia 31. * 495.
- - Memorialia in puncto juramenti &c. 284.
- - MSS. Antiſtitii Baſil. 342.
Actes de la paix de Nimegue 284.
- - - - - *d'Utrecht* 317.
Adami Relatio de Pacificat. Weſtphal. 26.
Adamus Bremensis 329.
Adelboldi vita Henrici S. 306. 544.
Adlzreiter annales Bojor. 529.
Adumbratio jurium Præfecturæ Hagenoenſis 263. ſ. 564.
Ægidius Aureæ Vallis geſta Epiſcoporum Leod. 480.
Æneæ Sylvii Epiſtolæ 312.
Agathias de Imperio Juſtiniani 275.
Alberici Chronicon 332. * 480. * 484. † 488.
Albertus Argent. 17. † 33. 37. &c. * 565. † 619. 620. 623. 624. 650. 654. * 695.
- - Stadiensis 11. * 63. † 189. 631.
Albini Chronicon Miſn. 23.
Albrecht *Hiſtorie von Hohenburg* † 208.
Alt, (*Baron*) *Hiſtoire des Suiſſes* 513.
Altingii Hiſt. Eccleſ. Palat. 182.
Andelo (Petr. de) de Imper. Rom. Germ. † 19. * 698.
Andler Corpus conſtitut. 316.
Annales Bertiniani 9.
- - Colmarienſes † 12. † 33. 37. &c. * 56. * 95. * 109. * 184. † 202. * 329. * 367. * 407. 622. † 624. 632. 634. 642. 653. 655.
- - Hildesheimenſes 188.
Anselme *Hiſt. Genealog. de France* 54. &c. † 473.
Argentin. Eccleſiæ agenda 342.
Aristotelis Opera 68.
Arnisæus de Republ. † 339.
Arnoldus Lubec. Chron. Slav. 551.
Arnulphus de ſchiſmate poſt Honorii II. deceſſum 461.
Arrêts notables du Conſeil d'Alſace 51. 60. &c.
Artopoei de redintegratione bonorum uxoris 326.
Athenæus 513.
Aventini Annales Bojor. 2.

B.

Baldericus Noviomensis Chron. Cameracenſe 480.
Balthasar (Jo. Frid.) de Judiciis S. R. Imp. 24.
Baluzii Capitularia 275. 630.
Bartholini Auſtrias ſive de Bello Norico 353. &c.
Baumann de conſortio Imperii &c. 502.
Bayoni Hiſtor. Mediani monaſt. 8. * 113. † 483. † 492. * 493.
Beck not. ad *Urſtiſii* Epit. Hiſt. Baſil. * 473.

a 2 Beck

Figure A.1 – Index auctorum



INDEX RERUM.

A.

Aab, *Aha*, *Ach* 79.
Abb' Reg. Scultetus Enſisheim. 67.
Aberzheim vicus 160.
Abresviller vicus 195.
Academiæ Alſatiæ 145. 182. 344. 346.
Aceti commercium 375.
Achen vicus 199.
Achenheim vicus 167. ſq. 684. Familia nobilis 634. Henricus 644.
Acidulæ Alſatiæ 89. 96. 351.
Adamsweiler vicus 198.
Adelbertsheim vicus 160.
Adelharteshoven 270.
Adelheidis Proprium 180.
S. Adelphi Eccleſia 224.
Adelsheim, familia nobilis 634. Georgius 209. Gœtzo 209. 573. Stephanus 209. Zeiſolf 209. poſſeſſiones 176. 209.
Adelshofen vicus deſtructus 270. 454.
Adern vicus 98.
Adgratiandi jus 325.
Adminiſtratorium jus 283.
Adolphi Imp. res geſtæ in Alſatia 109. 359. 371. 399. privilegia 369.
Adolsheim vicus deſtructus 454.
Adrazhofen villa 168. 270.
Advocati provinciales Alſatiæ 557. ſqq. 581. poſſeſſiones 253. 282.
Alsatiæ Ducatus 10. à Suevico ſeparatus 545. Duces 555.
Alſchweiler vicus deſtructus 85. ſq. 454. Familia nobilis 96. 634.
Alsgaudia Alſatiæ ademta 5.
Alſpach monaſterium 73. 415. 448. ſub advocatia Alſatiæ 282. poſſeſſiones 104. 106.
Alſtætter à Caldenburg 222.
Alſtingen vicus 240.
Altæ Silvæ monaſterium 408. advocatia 199. poſſeſſiones 270.
Altavilla (Jacobus) 155.
Altbeinheim capella 245.
Altdhau caſtrum 174.
Alteburg rupes 197.
Altenach vicus 38. Familia 43.
Altenau caſtellum 265. 431. 674.
Altenbach viculus 98.
Altenberg mons 270.
Altenburg caſtrum 462.
Altencaſtel caſtrum 108. Familia nobilis 634.
Altenheim v. Altheim.
Altenſtatt vicus 175. 387. 429.
Altenweiler villa deſtructa 153. 454. decimæ 301.
Altheimenſe Concilium 2. 539.
Altheim vicus deſtructus 78. 110. 115. 212. 215. 454.
Altglaßhütt villa 240.
Altkirch caſtrum 32. ſq. 37. 431. oppidum 37. 276. Dynaſtia 33. 37. ſqq.
Altmunſterol vicus 51.

Figure A.2 – Index rerum

Extraits des deux index de l'œuvre de Jean-Daniel Schoepflin [Source : http://bibliotheque-numerique.inha.fr/idurl/1/12532, p.804 et 813]

Annexe B

Les différents types d'interopérabilité

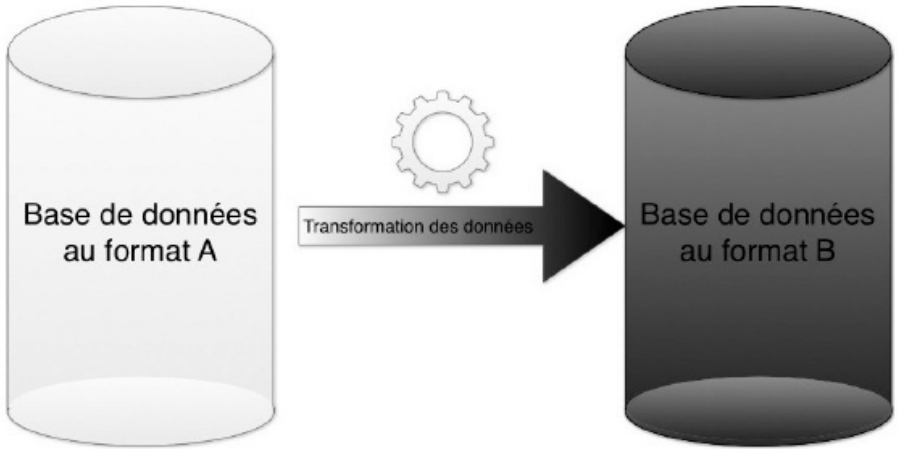


FIGURE B.1 – L'interopérabilité par conversion et copie [Source : BERMÈS (Emmanuelle), ISAAC (Antoine) et POUPEAU (Gautier), "2. Convergence et interopérabilité : vers le Web de données", *Bibliotheques* (, 2013), ISBN : 9782765414179 Publisher : Éditions du Cercle de la Librairie, p. 29-46, URL : `https://www-cairn-info.proxy.chartes.psl.eu/le-web-semantique-en-bibliotheque--9782765414179-page-29.htm` (visité le 01/08/2020)]

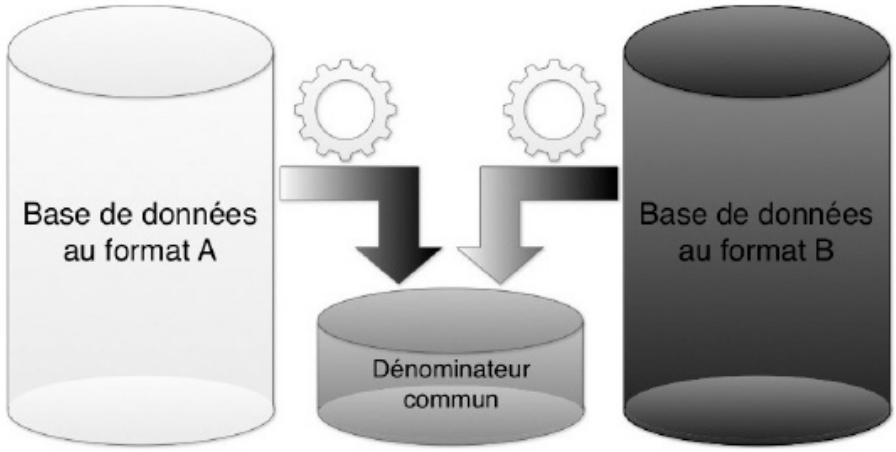


FIGURE B.2 – L'interopérabilité par le plus petit dénominateur commun [Source : BERMÈS (Emmanuelle), ISAAC (Antoine) et POUPEAU (Gautier), "2. Convergence et interopérabilité : vers le Web de données", *Bibliotheques* (, 2013), ISBN : 9782765414179 Publisher : Éditions du Cercle de la Librairie, p. 29-46, URL : `https://www-cairn-info.proxy.chartes.psl.eu/le-web-semantique-en-bibliotheque--9782765414179-page-29.htm` (visité le 01/08/2020)]

Annexe C

Les types de données présents dans les bases de données de l'INA et leur rôle

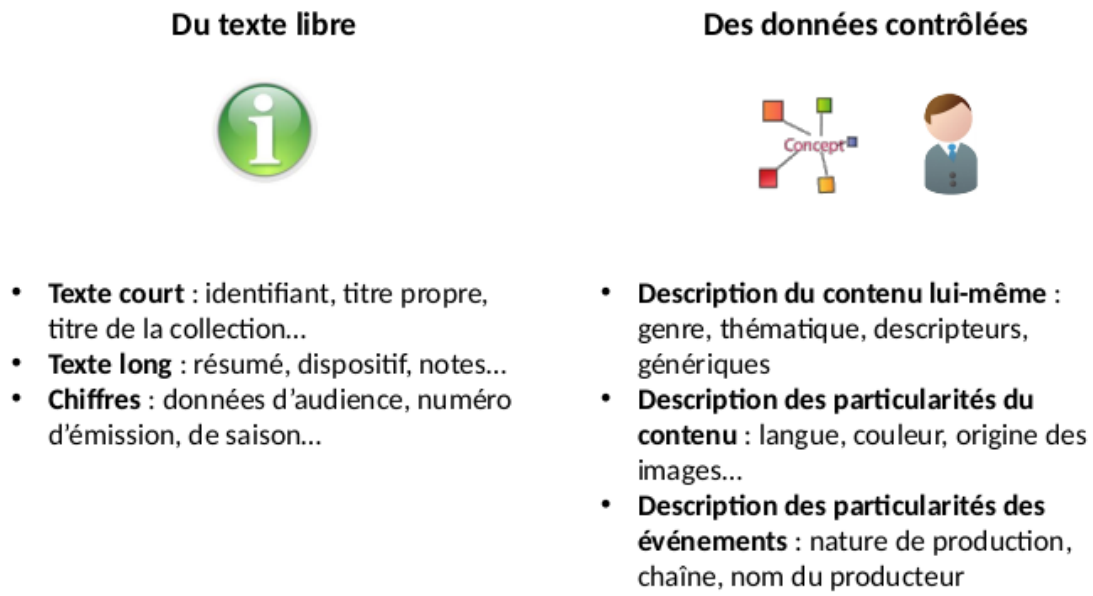


FIGURE C.1 – Les types de données présents dans les bases de données de l'INA [Source : ROCHE-DIORÉ (Axel), *Atelier transmission des connaissances*, 20 janv. 2020, p.6]

Annexe D

Le thésaurus de noms communs de l'INA

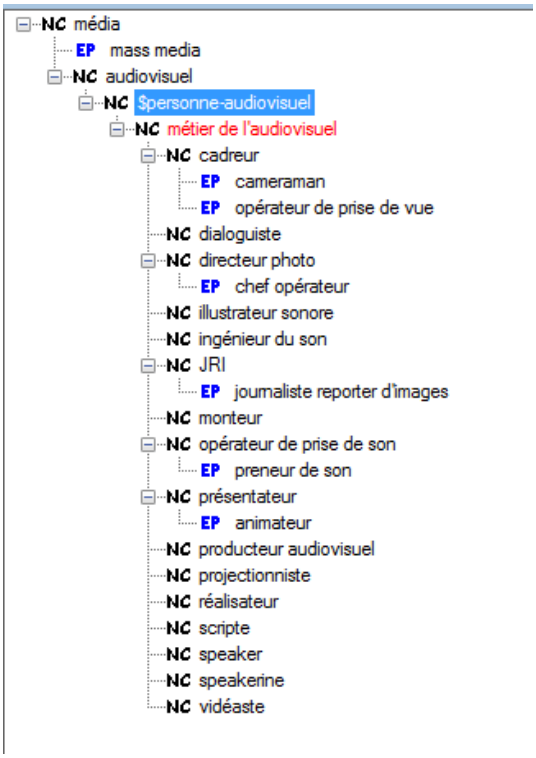


FIGURE D.1 – Extrait du thésaurus de noms communs de l'INA autour du terme « Cadreur »

Annexe E

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
jounaliste	
jounaliste audiovisuel	
jounaliste specialiste sant	
jounraliste	
jouornaliste	
jouranliste	
journalise	
journalisme	
journalitse	
journalliste	
journalsite	
jourrnaliste audiovisuel	

FIGURE E.1 – Résultat de l'alignement des journalistes avec le thésaurus des noms communs

Annexe F

Les captations directes réalisées par l'INA au titre du dépôt légal

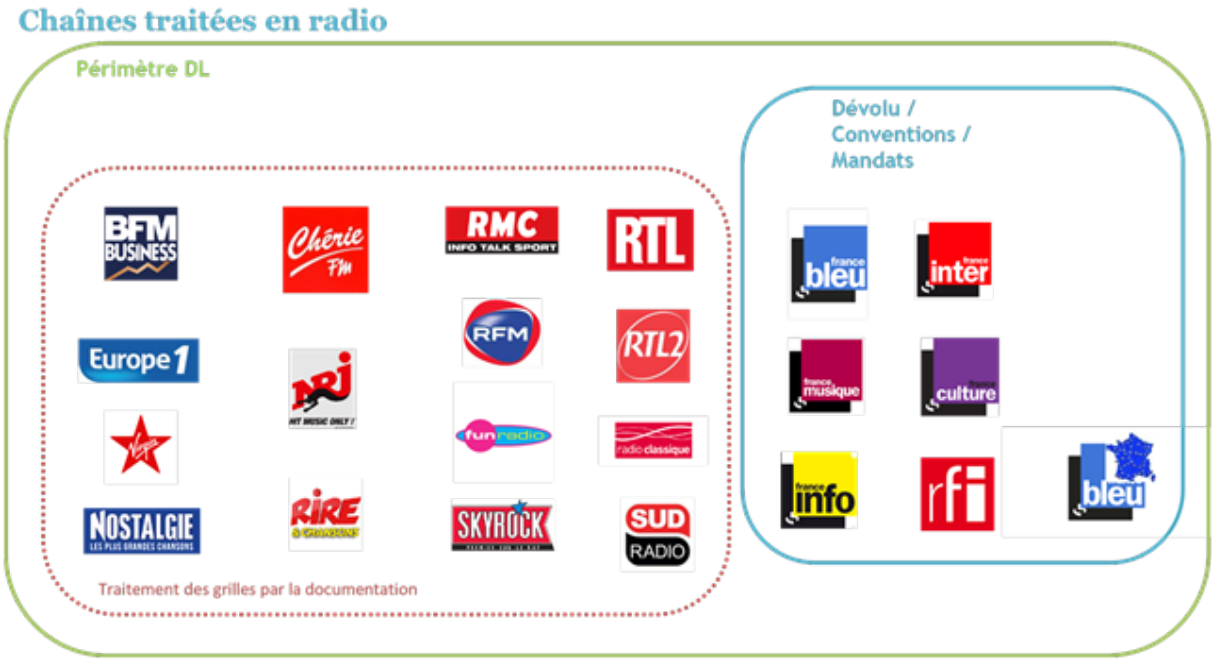


FIGURE F.1 – Stations de radio captées au titre du dépôt légal [Source : Communication électronique de l'entreprise « La collecte et le catalogage » du 26 mai 2020]

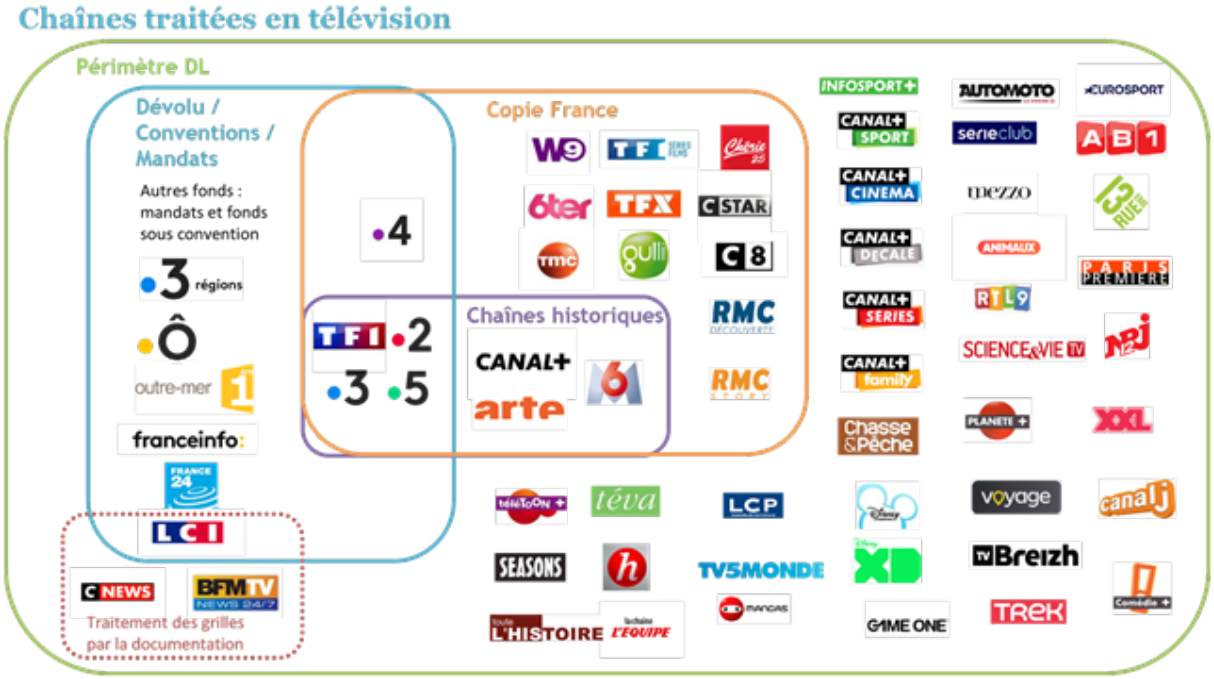


FIGURE F.2 – Chaînes de télévision captées au titre du dépôt légal [Source : Communication électronique de l'entreprise « La collecte et le catalogage » du 26 mai 2020]

Annexe G

Les fournisseurs externes de données de l'INA

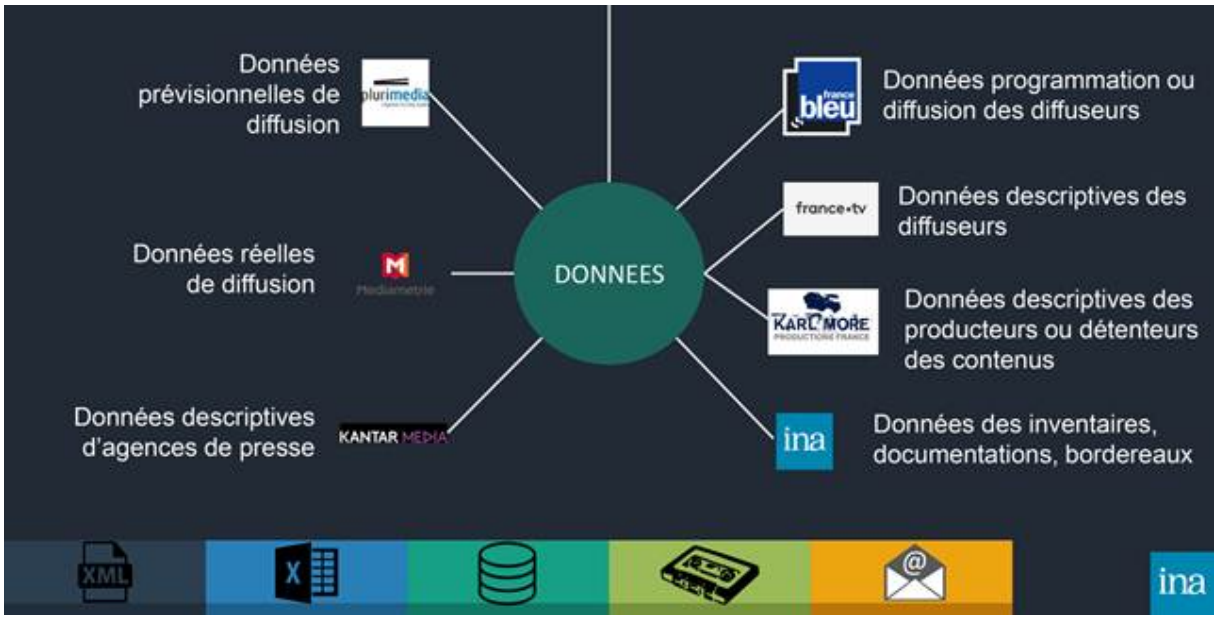


FIGURE G.1 – Les fournisseurs extérieurs de données de l'INA [Source : Communication électronique de l'entreprise « La collecte et le catalogage » du 26 mai 2020]

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 G.1 – L'apport des données de Médiamétrie dans la description effectuée au DL [Source : extrait de la base de données DLSAT de l'INA]

Index des noms de référentiels

Dewey, 25

FOAF, 26

GEMET, 25

LCSH, 5

RAMEAU, 5

Table des figures

1.1 Anneau de synonymie du terme « Prisonniers » de RAMEAU 8
1.2 Anneau de synonymie du terme « Prisoners » de LCSH 8

3.1 Arbre porphyrien de l'homme avec les seuls attributs analytiques 22
3.2 Arbre porphyrien prenant en compte les différences 23
3.3 Infinitude de l'arbre de Porphyre 23
3.4 Extrait de l'arborescence de l'index de Pline L'Ancien 24
3.5 Classification des vocabulaires selon leur complexité 26
3.6 Le modèle taxonomique 27
3.7 Le modèle du thésaurus simple 27
3.8 Le modèle polyhiérarchique 28
3.9 Application du modèle polyhiérarchique 28
3.10 Relation d'équivalence 29
3.11 Relation d'association 29
3.12 Relation de hiérarchie 30
3.13 Modélisation d'une vedette de thésaurus 30
3.14 Données d'exemple de notes qualité avec la fonction de Réalisateur 33
3.15 Données d'exemple de notes qualité avec la fonction de Réalisateur, après normalisation des fonctions 33
3.16 Données d'exemple de notes qualité sans fonctions 33
3.17 Utilisation des synonymes pour l'alignement du terme « Cuisinier » 35
3.18 Gestion de la polysémie dans l'alignement du terme « Cuisinier » 35

A.1 Index auctorum 57
A.2 Index rerum 57

B.1 L'interopérabilité par conversion et copie 59
B.2 L'interopérabilité par le plus petit dénominateur commun 60

C.1 Les types de données présents dans les bases de données de l'INA 61

D.1 Extrait du thésaurus de noms communs de l'INA 63

E.1 Résultat de l'alignement des journalistes avec le thésaurus des noms communs 65

F.1 Stations de radio captées au titre du dépôt légal 67
F.2 Chaînes de télévision captées au titre du dépôt légal 68

G.1 Les fournisseurs extérieurs de données de l'INA 69

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 **3**

1.1 Du langage libre au langage contrôlé : vers l'indexation 3

1.2 Une clé entre les données : les vocabulaires contrôlés 5

1.2.1 Contrôle de la forme des vedettes 6

1.2.2 Contrôle de la polysémie et de l'homographie 6

1.2.3 Contrôle de la synonymie 7

1.3 Une clé entre les jeux de données : l'interopérabilité par les fichiers d'autorité et les portails 7

1.3.1 La naissance des autorités par rétroconversion 9

1.3.2 Partager des vocabulaires : à la recherche de la meilleure interopérabilité 10

2 Les référentiels à l'INA **13**

2.1 De multiples fonds à décrire 14

2.1.1 Les archives professionnelles 14

2.1.2 Les fonds issus du dépôt légal 14

2.2 Un système documentaire pluriel répondant aux besoins 15

2.2.1 Les bases de données du dépôt légal (DL) 16

2.2.2 Les bases de données des archives professionnelles (DA) 17

2.2.3 La base de donnés juridique (DJ) 17

2.3 Multiplication des sources de données et des référentiels 18

3 L'arbre, un vocabulaire contrôlé hiérarchique 21

3.1 L'arbre de Porphyre : origines et influences 21

3.1.1 L'arbre de Porphyre 22

3.1.2 L'encyclopédisme (Antiquité - Moyen-Âge) : la recherche d'un arbre global mimant le monde réel 24

3.1.3 Influences : une diversité de référentiels hiérarchiques 25

3.2 Le *thésaurus*, vocabulaire contrôlé hiérarchique le plus fréquent 26

3.2.1 Types de structure 27

3.2.2 Relations entre les termes 28

3.2.3 Utiliser la précoordination pour les relations complexes 30

3.3 Passer du texte libre à un vocabulaire contrôlé : aligner des notes qualité et un thésaurus de noms communs 31

3.3.1 Contrôler du texte libre 31

3.3.2 Aligner les extractions en langage naturel avec un thésaurus de noms communs 34

3.3.3 Classer selon le thésaurus 35

II RELIER. Vers le partage de référentiels communs (début des années 2000 – milieu des années 2010) 37

4 Le web de données : une exposition commune des référentiels 39

5 Partager des structurations similaires de jeux de données par les classes et les propriétés : les ontologies, grammaires communes mais spécifiques 41

6 Relier ses données à Wikidata 43

III CENTRALISER. Le référentiel, clé de voûte et pivot (depuis le milieu des années 2010) 45

7 Les labyrinthes comme réseaux de données et de liens 47

8 Le Lac de données de l'INA : le référentiel au centre du modèle 49

9 Centraliser les référentiels de l'INA dans le Lac de données : l'exemple de l'alignement de deux référentiels de personnes physiques 51

Conclusion 53

Annexes **57**

A Les index de la Renaissance, termes contrôlés et classification alphabétique (les index de l'*Alsatia Illustrata* de Jean-Daniel Schoepflin) **57**

B Les différents types d'interopérabilité **59**

C Les types de données présents dans les bases de données de l'INA et leur rôle **61**

D Le thésaurus de noms communs de l'INA **63**

E Aligner les fonctions de « Journaliste » des notes qualité avec le thésaurus des noms communs de l'INA **65**

F Les captations directes réalisées par l'INA au titre du dépôt légal **67**

G Les fournisseurs externes de données de l'INA **69**

Index des noms de référentiels **71**

Table des figures **73**

Table des matières **75**