Les technologies du web sémantique et du record linkage au service de data.bnf.fr et du Linked Open Data culturel : étude sur les nouveaux paradigmes informationnels

Benjamin Duhamel

Benjamin Duhamel

Master 2 Gestion de l'Information et du Document pour l'Entreprise,
au sein du département des Sciences de l'Information et du Document.
Stage effectué du 01 avril au 29 août 2014 à la Bibliothèque Nationale de France,
au sein du département de l'Information Bibliographique et Numérique.
Mémoire soutenu le 10 septembre 2014 à l'université de Lille 3

Les technologies du web sémantique et du *record linkage* au service de data.bnf.fr et du *Linked Open Data* culturel : Étude sur les nouveaux paradigmes informationnels

Sous la direction de Gildas Illien (tuteur professionnel, BnF)
et Gérald Kembellec (tuteur universitaire, Lille3)

Année universitaire 2013 -2014

Mes remerciements à Agnès Simon et Sébastien Peyrard pour leur soutien constant, leur générosité et leur bienveillance,

à Gildas Illien, Françoise Bourdon et Jérôme Villeminoz pour m'avoir accueilli dans leur équipes dans des conditions optimales,

à l'ensemble des collègues très chaleureux qu'il m'a été offert de rencontrer à la BnF.

Mes remerciements à Gérald Kembellec pour m'avoir guidé efficacement dans l'approche et la réalisation de ce stage et du mémoire.

Terminant un cycle démarré il y a 5 ans, j'ai une pensée également pour les nombreuses personnes qui m'ont aidé à parcourir ce chemin : Stéphane Tywoniuk, Patrick Verrougstraete, Christine van Lancker, Béatrice Micheau, Monique Cordonnier, Joachim Schöpfel, ainsi que ceux dont j'ai oublié de citer le nom mais que je garde en mémoire.

Enfin, mes pensées vont également naturellement à ma marraine et mon oncle, mes parents ainsi que celle qui partage ma vie.

Résumé

Les technologies du web sémantique et du *record linkage* changent profondément les pratiques et réseaux habituels des professionnels du patrimoine culturel. De nouvelles modalités de production et de diffusion des métadonnées émergent dans un nouvel écosystème, le *Linked Open data* culturel, où les partenaires comme les usages évoluent. On observe une porosité des frontières traditionnelles entre les acteurs du monde de la culture (bibliothèques, musées, archives), mais aussi entre ces derniers et les nouveaux acteurs du web (e.g. Wikipédia). Ce nouveau paysage de métadonnées ouvertes et liées ouvre des perspectives de rationalisation des coûts et de description efficiente et distribuée des ressources. Un nouveau profil d'usager apparaît, le réutilisateur de données, qui développe de nouveaux services en agrégeant des jeux de données auparavant distincts. L'usager plus traditionnel bénéficiera également de l'interopérabilité entre les bases de données, ce qui permettra d'améliorer son expérience de navigation et de recherche d'information. Il participera lui-même à l'enrichissement de la production grâce aux promesses du web social-sémantique. Cet environnement en mutation nécessite une gouvernance redéfinie afin d'établir la confiance entre ses différents acteurs et assurer son bon fonctionnement à long terme. Cela ouvre aux institutions culturelles l'opportunité stratégique de consolider leur place sur le web de données ; une des conditions d'un tel succès est d'incorporer de nouvelles manières de travailler à leurs activités traditionnelles, en trouvant notamment une complémentarité efficiente entre l'humain et les technologies algorithmiques de traitement de jeux de données en masse.

Mots-clefs : données liées ouvertes, web sémantique, algorithmes, alignements, bibliothèques, culture, métadonnées, vocabulaires d'autorité, gouvernance

Abstract

Semantic web and record linkage technologies cause very deep changes in professional networks and practices in the cultural heritage domain. New methods for creating and disseminating metadata surface on a new ecosystem, the cultural Linked Open Data, where partners and users evolve. The traditional boundaries tend to fade between the traditional stakeholders from the cultural area (archives, libraries and museums) , but also between those stakeholders and new actors from the web like Wikipedia. This new linked open metadata space opens up opportunities for rationalizing costs and for efficiently distributing the resource description task. A new user profile emerges with data re-users, who build new services by aggregating previously distinct datasets. Traditional users also benefit from such interoperability across databases as it enables enhanced browsing experience and information retrieval. They can also contribute to enrich resources thanks to social semantic web promises. This mutating environment requires a redefined governance to build trust between all its stakeholders and ensure its sustainability. This opens a strategic opportunity for cultural institutions to consolidate their role in the web of data. To succeed in this, it is crucial that they incorporate new worflows to their traditional activities, and find, among other things, an efficient balance between human procedures and massive algorithmic processes.

Keywords: linked open data, semantic web, record linkage, algorithms, alignments, libraries, cultural heritage, metadata, authorities, value vocabularies, governance

Sommaire

Introduction générale

Dans ce mémoire de fin d'études de Master 2 GIDE en Sciences de l'Information, vous trouverez deux rapports : d'une part un rapport de stage centré sur les missions de gouvernance des métadonnées de la BnF avec les technologies du web sémantique et du *record linkage*, dans le cadre du projet data.bnf.fr. D'autre part un rapport de recherche centré sur la littérature scientifique au sujet des enjeux technologiques, professionnels et stratégiques liés à l'alignement de vocabulaires d'autorités sur le web de données culturel.

Ces deux parties permettront finalement d'aborder une même problématique autour du développement d'un nouvel écosystème *Linked Open Data* dans le domaine culturel, sous deux angles différents : que ce soit du point de vue d'une institution, acteur majeur de cet environnement en mouvement, la Bibliothèque National de France, que d'un point de vue plus macro, visant à mettre à jour les tendances globales de ce changement de paradigme.

Partie I. Rapport de stage : Gouvernance des métadonnées de la BnF avec les technologies du web sémantique et du record linkage, dans le cadre du projet data.bnf.fr

A) Introduction

A.1 Le projet data.bnf.fr à la Bibliothèque nationale de France (BnF) : pour une nouvelle gouvernance des données sur le web de données

La Bibliothèque nationale de France a pour mission la diffusion de la bibliographie française issue du dépôt légal[1], institué en 1537 par François Ier, et l'offre de sa « consultation à distance en utilisant les technologies les plus modernes de transmission des données » (décret de 1994 portant création de la BnF).

Au sein de la BnF, le département de l'Information Bibliographique et Numérique (IBN), prolongement de l'ancienne « Agence Bibliographique Nationale », assure les fonctions de coordination du catalogage et d'établissement des règles de descriptions normalisées nécessaires à la production et la diffusion de la bibliographie nationale. Son travail porte sur la structure des métadonnées descriptives de nature bibliographique (description de documents) et d'autorité : points d'accès personnes physiques, organisations, titres, classification décimale Dewey, sujets RAMEAU, notices géographiques, marques. Au sein du département de l'IBN, le Service Prospective et services documentaires suit en amont et en aval du Service Coordination et administration des métadonnées (production des données), les évolutions des normes et modèles de données associées aux usages, ainsi que la diffusion des produits et services bibliographiques vers l'extérieur.

Le projet data.bnf.fr[2], en ligne depuis juillet 2011, fut au départ une preuve de concept, pour être devenu aujourd'hui un véritable service, ayant pour objectif de regrouper les informations des différents catalogues et bases de la BnF (catalogue général[3], catalogue archives et manuscrits[4], reliures.bnf.fr[5], bibliographie des éditions parisiennes du 16e siècle[6], dépôt légal du Web[7] et, à terme, les expositions virtuelles[8]) et de sa bibliothèque numérique Gallica[9], pour les rendre plus visibles et plus accessibles sur le Web, ce qui en fait un pivot documentaire[10].

---

[1] « *Le dépôt légal est l'obligation pour tout éditeur, imprimeur, producteur, importateur de déposer chaque document qu'il édite, imprime, produit ou importe en France à la BnF* ».
[2] Voir http://data.bnf.fr/
[3] Voir http://catalogue.bnf.fr/
[4] Voir http://archivesetmanuscrits.bnf.fr/
[5] Voir http://reliures.bnf.fr/
[6] Voir http://bp16.bnf.fr/
[7] Voir http://www.bnf.fr/fr/professionnels/depot_legal/a.dl_sites_web_mod.html
[8] Voir http://expositions.bnf.fr/
[9] Voir http://gallica.bnf.fr/
[10] Wenz Romain. « Data.bnf.fr : au-delà des silos ». *Approches documentaires : priorité aux contenus*.

Le projet data.bnf.fr suit les évolutions en cours des principes de catalogage établis en 2008 par l'IFLA[11] visant à mieux répondre aux besoins informationnels des utilisateurs : trouver, identifier, sélectionner, se procurer, naviguer. C'est ainsi que le modèle de données de data.bnf.fr suit les principes de la modélisation fonctionnelle FRBR (*Functional Requirements for Bibliographic Records*)[12] : le site data.bnf.fr a pour ambition de guider les utilisateurs à travers les ressources de la BnF, en publiant des pages de référence sur les auteurs, les œuvres, les thèmes ou les lieux.

Les données sont exposées et mises à disposition selon les standards du web sémantique, en utilisant le formalisme RDF (*Ressource Description Framework*), et sont également librement réutilisables car sous la licence ouverte de l'État[13]. Ainsi est encouragée et facilitée la réutilisation des données, que ce soit au sein de la communauté des bibliothèques, mais aussi en dehors de celle-ci, dans l'esprit du web de données qui abolit les frontières métiers traditionnelles[14]. Dans l'environnement du Linked Open Data, la BnF, via data.bnf.fr, lie ses données avec d'autres, présentes sur le web de données, dans un écosystème ouvert de métadonnées liables[15].

Les notices du catalogue sont ainsi atomisées pour devenir des données indépendantes de leur contexte d'origine ou de leur mise en application dans un catalogue de bibliothèque. Ces données sont liées à d'autres jeux de données sur le web : les données d'autorités « Auteurs » de la BnF (personnes physiques et organisations) sont liées dans data.bnf.fr, c'est à dire alignées, avec celles de VIAF[16] (Virtual International Authority File), ISNI[17] (International Standard Name Identifier), IdRef[18] (Identifiants et Référentiels pour l'enseignement supérieur, ABES[19]) et DBpedia[20]. Le référentiel d'autorités matière RAMEAU[21], encyclopédique et contrôlé, est aligné dans data.bnf.fr avec d'autres référentiels sujets tels qu'AGROVOC[22], le Thésaurus W[23] (pour la description et l'indexation des archives locales anciennes, modernes et

[11] Fédération internationale des associations de bibliothécaires et d'institutions
[12] Voir http://www.bnf.fr/fr/professionnels/modelisation_ontologies/a.modele_FRBR.html
[13] Voir http://www.etalab.gouv.fr/pages/licence-ouverte-open-licence-5899923.html
[14] Groupe d'incubation W3C. Rapport final du groupe d'incubation « Bibliothèques et web de Données ». URL : http://mediatheque.cite-musique.fr/MediaComposite/ARTICLES/W3C/XGR-lld-fr.html
[15] Ibidem.
[16] Voir http://viaf.org/
[17] Voir http://www.isni.org/
[18] Voir http://www.idref.fr/autorites/autorites.html
[19] Agence bibliographique de l'enseignement supérieur
[20] Voir http://dbpedia.org/About
[21] Répertoire d'autorité-matière encyclopédique et alphabétique unifié
[22] Voir http://aims.fao.org/standards/agrovoc/about
[23] Voir http://www.archivesdefrance.culture.gouv.fr/gerer/classement/normes-outils/thesaurus/

contemporaines), les autorités matière de la Bibliothèque nationale Allemande[24] (DnB[25]) et de la Bibliothèque du Congrès[26] (LOC[27]), la classification décimale Dewey[28] ou encore Geonames[29] dans le cas des sujets géographiques.

Les données sont récupérées par des tiers pour des usages qui n'étaient pas prévus au préalable : de par leurs caractéristiques, ces données étant contrôlées, sourcées, datées et produites par une institution publique faisant autorité, elles inspirent la confiance et trouvent une vraie valeur ajoutée sur le web de données. Elles sont utilisées par exemple dans le monde de la recherche avec le projet Isidore[30], ou dans des applications innovantes comme l'espace numérique de travail éducatif Abuledu[31].

Dans ce contexte, il m'était confié plusieurs missions dans le cadre de ce stage. Premièrement, la réalisation d'alignements et de traitements de données pour data.bnf.fr. Il s'agissait notamment d'améliorer l'algorithme d'alignement entre les deux référentiels géographiques internes de la BnF : les « noms géographiques »[32] (GEO) du département des Cartes et Plans et les « Rameau noms géographiques »[33] (RAM NG). L'objectif est de créer et d'automatiser une passerelle entre ces deux référentiels conceptuellement proches et rapprochables mais dont les « *records* » sont en silos.

J'ai également participé à la création manuelle de liens entre données. En effet, le web de données, même s'il vise à automatiser beaucoup de processus, ne rentre pas non plus en contradiction avec la réalisation d'annotations descriptives manuelles. C'est ainsi que j'ai contribué à l'indexation des expositions virtuelles[34] avec des sujets d'autorité (auteurs, titres, RAMEAU), afin qu'elles puissent être retrouvées dans data.bnf.fr (par exemple, depuis la page de Katsushika, Hokusai (1760-1849), on retrouvera alors le lien vers son exposition virtuelle BnF dédiée[35]). Je me suis donc placé tant du point du vue de la production manuelle de données que de leur traitement automatique a posteriori.

[24] Voir http://d-nb.info/standards/elementset/gnd#
[25] Deutsche Nationalbibliothek
[26] Voir http://id.loc.gov/authorities/subjects.html
[27] Library of Congress
[28] Voir http://dewey.info/
[29] Voir http://www.geonames.org/
[30] Voir http://www.rechercheisidore.fr/referentiels
[31] Voir http://www.abuledu.org/
[32] Voir http://www.bnf.fr/fr/professionnels/autorites_bnf/s.noms_geographiques_bnf.html
[33] Voir http://guiderameau.bnf.fr/html/rameau_0894.html
[34] Voir http://expositions.bnf.fr/
[35] Voir http://expositions.bnf.fr/lamer/feuille/index10.htm

J'ai contribué à la mise en place du SPARQL *endpoint*[36] de data.bnf.fr, outil permettant d'interroger finement des triplets RDF. Nous avons mis en place des tests ainsi que des profilages de requêtes et de jeux de données pour des cas de réutilisation.

Ce travail a également permis de parcourir et d'explorer efficacement le modèle de données et a ainsi été mené conjointement à la mise à jour de la documentation du modèle de données RDF de data.bnf.fr[37] ainsi qu'à la proposition de modification de certaines propriétés. Le profilage de jeux de données pour des cas de réutilisation comme la documentation du modèle de données trouvent leur importance dans la communication et la présentation du projet data.bnf.fr auprès de ses réutilisateurs potentiels.

Enfin, le département de l'IBN a déployé une phase de tests précédant la réinjection dans le catalogue des alignements calculés automatiquement dans data.bnf.fr entre des notices bibliographiques et leurs « autorités titres » correspondantes. En effet, le MARC est toujours le format de production des données bibliographiques de la BnF et la question d'y réinjecter des liens pertinents créés dans data.bnf.fr se pose donc. En l'occurrence, il s'agit ici des liens permettant de « FRBRiser le catalogue » en raccrochant des publications (les manifestations) à leur autorité (l'œuvre).

A.2 Problématisation

A.2.1 data.bnf.fr, un service qui permet de repenser le système d'information et les méthodes de travail, en adéquation avec les évolutions des technologies et pratiques sur le web

Avec l'utilisation des technologies et standards du web sémantique, c'est l'ensemble de l'administration des métadonnées qui est repensée pour leur diffusion et leur usage sur le web. Au niveau du modèle « métier », data.bnf.fr se départit des présentations habituelles de l'information bibliographique sous la forme de l'International Standard Bibliographic Description[38] (ISBD) pour se fonder sur le modèle conceptuel FRBR, centré sur les pratiques des utilisateurs sur le web (voir le schéma global en **annexe I**). L'élément central est que « *les FRBR organisent les différentes composantes de la description bibliographique (les autorités, les accès sujet et les informations sur le document proprement dites) en trois groupes d'entités reliées ensemble par des relations* »[39].

[36] Voir http://data.bnf.fr/sparql
[37] Voir http://data.bnf.fr/images/graphe_complet.pdf
[38] Voir http://www.bnf.fr/fr/professionnels/normes_catalogage/a.normes_isbd_presentation.html
[39] Figoblog, le blog d'Emmanuelle Bermès. *Les FRBR, qu'est-ce que c'est ?* http://www.figoblog.org/document594.php

Au niveau du « document » et/ou de l'œuvre donc, ces notions sont représentées graduellement de la plus conceptuelle à la plus matérielle, en « œuvre », « expression », « manifestation » et finalement « item ». Cette modélisation met en avant les données d'autorité et les liens qu'elles entretiennent entre elles, alors que la notion de document intervient alors dans un temps ultérieur. Ceci est intéressant car jusqu'à présent la pratique était inverse : celle-ci partait du document pour créer utilement des autorités pour les accès. Avec les FRBR, c'est l'autorité qui devient l'élément central auquel on rattache des documents ou d'autres autorités.

Pour ce qui est des formats de métadonnées d'origine, ils varient selon les spécificités et besoins « métiers » et sont complétés par un cadre normatif qui permet d'indiquer comment les utiliser. Par exemple, les formats MARC (MAchine-Readable Cataloging) sont utilisés pour les produits bibliographiques afin d'encoder des descriptions bibliographiques[40], alors que l'EAD (Encoding Archival Description) est utilisé pour encoder des descriptions de manuscrits ou de documents d'archives[41].

La diffusion de ces données sur le web sémantique a conduit tout d'abord à la mise en correspondance (*mapping)* de l'information structurée encodée dans ces formats informatiques « métiers » avec une structure cible, utilisant désormais les vocabulaires du web sémantique, mélangeant des schémas plus généralistes tels que Friend Of A Friend[42] (FOAF) ou Dublin Core[43], ainsi que « métiers » comme les vocabulaires de Ressource Description and Access[44] (RDA) (Voir un exemple en **annexe II**).

Les standards du web sémantique normalisés par le W3C[45] : Resource Description Framework (RDF) et Uniform Ressource Identifier (URI), permettent d'exprimer et migrer ces descriptions dans un formalisme adéquat où la notice n'existe plus, où chaque ressource est identifiée par une URI et où la grammaire du triplet « sujet – prédicat – objet » de RDF permet d'exprimer et de multiplier les annotations de ressources sans restrictions de cardinalité (voir un exemple en **annexe III**) dans un « *formalisme (…) à même d'assurer l'interopérabilité des données sur le web* »[46]. Notons que le standard RDF « *s'impose de plus*

[40] Dont le cadre normatif générique est l'ISBD, lui-même complété par les normes nationales (e.g.la norme AACR issue de la tradition anglo saxonne)
[41] Dont Le cadre normatif générique utilisé en général est l'ISAD(G)
[42] Voir http://www.foaf-project.org/
[43] Voir http://dublincore.org/documents/dcmi-terms/
[44] Voir http://rdvocab.info/
[45] World Wide Web Consortium : http://www.w3.org/
[46] Bermès Emmanuelle. *Le Web sémantique en bibliothèque*. P. 43

en plus comme un outil d'unification conceptuelle et technique »[47] pour l'organisation des connaissances, afin d'identifier, de décrire et de mettre en relation « *tous les types de ressources, réelles ou virtuelles, physiques ou abstraites* »[48].

Avec cette approche, « *les entités, les relations et leurs attributs sont explicités et formalisés* »[49], ce qui a pour effet majeur de pouvoir typer les liens de manière à ce qu'ils soient exploitables par une machine. RDF se pose aussi comme un modèle « *ouvert* » et « *extensible* » qui ne rend pas une description « *prisonnière d'un schéma de métadonnées* »[50]. Enfin, pour effectuer les liens dans ce modèle de graphe, il est nécessaire de procéder à des alignements entre vocabulaires. Ici nous ne parlons pas d'alignements d'ontologies, mais d'alignements entre vocabulaires contrôlés ou référentiels d'autorités (thésaurus, systèmes de classifications, listes d'autorités) exposés sur le web de données[51]. La plus-value du web sémantique, est que l'on peut aligner « *deux jeux de données même si ceux-ci sont exprimés dans un modèle différent (...) pourvus qu'* [ils] *aient un point de contact* »[52].

A.2.2 Médiation du modèle de données RDF et de requêtes SPARQL en vue des réutilisateurs et création de liens manuels ou automatiques pour data.bnf.fr

Les données du site data.bnf.fr sont interrogeables depuis le mois d'août 2014 avec un requêteur SPARQL. Cet outil permettant de requêter finement les données exposées en RDF, a nécessité au préalable des opérations de tests pour en vérifier la pertinence des résultats ainsi que la robustesse du système. Egalement, c'est la question de l'usage de cet outil qui s'est posée, et en conséquence le constat d'un besoin de médiation vers les réutilisateurs potentiels, que ce soit au niveau du modèle de données RDF que du profilage de jeux de données avec SPARQL *endpoint*.

Si cet outil demande une sensibilité informatique avérée pour sa prise en main, le profilage de jeux de données avec des requêtes types et les résultats associés est une étape pour rendre accessible cet outil aux bibliothécaires comme aux professionnels d'autres communautés, de l'information-documentation ou d'autres horizons. Surtout, il permet à des développeurs d'appréhender plus facilement le modèle de données de data.bnf.fr au niveau métier (e.g.

[47] Vatant Bernard. « *Des métadonnées à la description de ressources : les langages du Web Sémantique* » p164
[48] Ibidem.
[49] Dalbin Sylvie. « Métadonnées et normalisation ». In Métadonnées : mutations et perspectives. P. 16
[50] Vatant Bernard. « *Des métadonnées à la description de ressources : les langages du Web sémantique* ». p. 169
[51] e.g. LCSH, AGROVOC, VIAF, Dewey, GeoNames, IdRef, INSEE, IGN
[52] Bermès Emmanuelle. *Le Web sémantique en bibliothèque*.

FRBR, distinction entre le concept et la personne à proprement parler, etc.). Il est aussi pédagogique de représenter le modèle de données RDF, selon des choix graphiques associant vue générale et zooms, afin d'exprimer efficacement et de manière aussi bien précise que globale un modèle de plus en plus complexe, intégrant progressivement de nouveaux type de données : spectacles, périodiques, lieux et dates, par exemple.

Ces travaux visent à aider les réutilisateurs, quels que soient leurs besoins informationnels et leurs intérêts, qu'ils soient plutôt développeurs ou fonctionnels, maîtres d'œuvre ou maîtres d'ouvrage, ou encore chercheurs.

Le projet data.bnf.fr s'inscrivant dans une démarche visant à faire évoluer les données du catalogue vers les modèles de données FRBR, différentes méthodes reposant sur la création de liens permettent de parvenir à cet objectif : en utilisant des algorithmes d'alignements de référentiels permettant de calculer ces liens ou en indexant manuellement les ressources.

L'alignement que nous avons étudié est un rapprochement entre deux jeux de données internes de la BnF : il s'agit de l'alignement entre les noms géographiques utilisés par le département des cartes et plans pour indexer leurs documents cartographiques, et les notices Rameau de type noms géographiques utilisées pour l'indexation sujet dans le cadre de la bibliographie nationale.

Ces deux référentiels existent conjointement, pour des raisons historiques et de contraintes spécifiques à la production des données, et l'enjeu avec les technologies du web sémantique est de les lier pour regrouper les ressources utilisant les instances identiques de ces référentiels. L'alignement réalisé permet par exemple d'avoir un accès unifié aux ressources de la ville de Paris, indexées dans l'un ou l'autre référentiel, depuis une même page de data.bnf.fr : http://data.bnf.fr/lieu/paris__france_/. Nous le verrons, une première phase d'alignement avait été déjà réalisée et l'objectif était d'améliorer l'algorithme d'alignement avec des règles métiers, afin d'aligner plus de « *records* » ensemble.

Par ailleurs, le travail d'annotation des expositions virtuelles avec des autorités BnF évoqué précédemment, est également important d'un point de vue stratégique car il rappelle aussi que le web de données n'est pas réservé uniquement à l'automatisme des machines : sa plus-value est de replacer toutes les contributions permettant d'enrichir sémantiquement des ressources, par des humains ou des machines, dans un espace d'information ouvert. Les enjeux liés à la question de la réinjection dans le catalogue des liens créés dans data.bnf.fr, permettant de « FRBRiser » la production, ont également occupés une part de mes missions.

Ainsi le département de l'IBN m'a inclus à son panel de testeurs, internes à la BnF, chargé d'effectuer des contrôles au cœur des données, sur la pertinence des alignements calculés entre des « autorités titres » (œuvres) et des notices bibliographiques (manifestations).
Nous avons pu identifier les erreurs commises par des processus d'alignements automatiques, en chercher la cause et essayer de trouver des solutions pour améliorer les algorithmes. Le catalogue est toujours l'outil de production des métadonnées de la BnF alors que data.bnf.fr est un outil de diffusion. Le sujet de la qualité des données est un enjeu crucial pour la BnF qui a une réputation de fiabilité. Les problématiques d'erreurs occasionnées par les alignements automatiques ont conduit le département de l'IBN à reconsidérer ces biais créés par les alignements automatiques, et à chercher des moyens de les résoudre, d'autant plus lorsqu'il s'agit de la réinjection de ces calculs dans les données sources.

A.3 Présentation du plan

Nous présenterons dans la partie B les réalisations effectuées dans le cadre des missions de stage. Nous étudierons dans une première partie (B.1) le sujet de la médiation du système d'information de data.bnf.fr en direction de ses nouveaux usages et usagers à travers la documentation du modèle de données RDF et de l'implémentation des FRBR, ainsi que du profilage de jeux de données avec le SPARQL *endpoint*. Dans une seconde partie (B.2), nous analyserons la question de la création des liens entre les ressources pour data.bnf.fr dans le cadre d'une approche *Linked Entreprise Data* permettant de « FRBRiser » les données. Ce travail nous amènera à nous intéresser aux nouvelles manières de travailler pour annoter et mettre en relation des ressources et aux enjeux professionnels qui en découlent, liés aux rapports hommes / machines. Nous traiterons d'abord le cas des alignements Géo – Rameau avec l'utilisation d'algorithmes d'alignements automatiques. Puis nous présenterons un cas d'indexation manuelle dans le cadre de la création de liens avec les expositions virtuelles. Enfin nous présenterons le travail d'affinage des algorithmes d'alignement entre les notices bibliographiques et les notices d'autorité titres, dans la perspective de la réinjection de ces liens calculés dans les données sources du catalogue.
Dans la partie C, nous discuterons le travail réalisé, que ce soit par rapport à la méthodologie utilisée ou à ses résultats.

B) Gouvernance des données catalographiques de la BnF avec les technologies du web sémantique et du record linkage[53]

B.1 Médiation du système d'information de data.bnf.fr en direction des nouveaux usages et usagers

B.1.1 Diffusion des données selon les principes et règles des FRBR et du Linked Open Data : un besoin de médiation

Le projet data.bnf.fr se fonde sur le modèle conceptuel des FRBR, utilise le formalisme de données RDF, la *« grammaire universelle des machines »*[54], les URIs pour identifier chaque chose, ou « *ressource* »[55], et les ontologies et vocabulaires du Web sémantique pour exprimer les entités et leurs propriétés. La notice et son schéma de métadonnées laisse place aux triplets, ou « *déclarations* », dont « *la ressource est le sujet* (…), *la propriété en est le prédicat, et la valeur de la propriété [en] est l'objet* »[56]. Au niveau de l'encodage informatique, le document laisse place aux données (une des sérialisation des données du *triplestore*[57] data.bnf.fr est le RDF/XML).

Le modèle de données RDF de data.bnf.fr (voir la documentation sur le site[58]) reprend donc les trois grands groupes spécifiés dans le rapport de l'IFLA[59] pour les FRBR : tout d'abord le groupe 1 avec les entités Œuvre, Expression et Manifestation, permettant de représenter « *les différents aspects de ce qu'un utilisateur peut trouver dans les produits d'une activité intellectuelle ou artistique* »[60]. Ensuite, les entités du groupe 2, représentant la mention de « *responsabilité du contenu intellectuel ou artistique* », sont divisées dans le modèle de data.bnf.fr en deux entités : Personne et Organisation. Enfin, le groupe 3, représentant le *« nouvel ensemble d'entités qui constituent les sujets d'œuvres »* est caractérisé dans data.bnf.fr par les entités Concept (au sens des sujets Rameau) et Lieu.

---

[53] Voir http://en.wikipedia.org/wiki/Record_linkage

[54] Poupeau, Gautier. « RDF,la grammaire universelle des machines ». in *Enjeux et technologies : des données au sens*. Documentaliste-Sciences de l'information

[55] « *les ressources sont un concept de base sur le web sémantique : tout ce à quoi on peut se référer est considéré comme une ressource par exemple une page web, une image, une vidéo, mais aussi une personne, un lieu, un dispositif, un évènement, une organisation, un produit ou un service (…) tout ce qui peut être identifié par un URI peut être considéré comme une ressource* ». [55] def. de Gandon Fabien, Faron-zucker Catherine, Corby Olivier. Le Web sémantique : Comment lier les données et les schémas sur le web ? p.28

[56] Gandon Fabien, Faron-zucker Catherine, Corby Olivier. Le Web sémantique : Comment lier les données et les schémas sur le web ? p.29

[57] Voir http://fr.wikipedia.org/wiki/Triplestore

[58] Modèle de données RDF data.bnf.fr synthétisé : http://data.bnf.fr/semanticweb#Ancre3 et général : http://data.bnf.fr/images/graphe_complet.pdf

[59] Groupe de travail IFLA. Fonctionnalités requises des notices bibliographiques : rapport final : http://www.bnf.fr/documents/frbr_rapport_final.pdf

[60] Le niveau Item n'est représenté que lorsqu'il y a un exemplaire numérisé dans Gallica, la fonction de localisation d'exemplaire restant déléguée au catalogue

Il faut noter que dans ce modèle, deux éléments supplémentaires propres au modèle de data.bnf.fr sont à expliciter : d'une part, la présence de l'entité Date, qui ne renvoie pas à un référentiel dans le catalogue ni à une entité des FRBR ; d'autre part et surtout, la présence de l'entité Concept dont le sens ici est plus large que celui évoqué précédemment dans le cas des sujets Rameau. En effet, dans data.bnf.fr, on distingue la notice catalographique sur un « objet » (e.g. la notice d'autorité de la BnF sur Victor Hugo) de l' « objet » réel identifié virtuellement sur le web de données, « ex : La personne de Victor Hugo). Le premier cas est exprimé dans data.bnf.fr comme un Concept alors que le second cas est exprimé comme un Auteur, ces deux entités étant par ailleurs reliées.

Ceci posé, le modèle de data.bnf.fr rajoute ensuite la couche des ontologies et vocabulaires du web sémantique pour exprimer les entités par des « classes », les attributs d'entités et les relations entre entités par des propriétés. Ainsi pour les auteurs, l'entité Personne est qualifiée comme étant de la classe « Person » et l'entité Organisation de la classe « Organization », avec le vocabulaire RDF Friend Of A Friend (FOAF)[61]. L'entité Œuvre est caractérisée par la classe « Work » avec le vocabulaire frbr-rda : une relation entre ces deux entités (et donc ces deux classes) par exemple la responsabilité intellectuelle d'une personne sur une oeuvre, peut s'exprimer avec la propriété « creator » du schéma de métadonnées Dublin Core (DC)[62]. Enfin, les attributs d'une entité ou d'une classe sont exprimés en ayant recours à différentes terminologies reprenant des schémas de métadonnées. Ainsi, on réutilisera pour une personne physique les éléments de description « name », « gender » et « birthday » de FOAF et les éléments de description « dateOfBirth », « placeOfBirth » et « biographicalInformation » du vocabulaire rdagroup2elements[63]. On voit ainsi que dans le modèle nous pouvons mélanger des vocabulaires RDF différents pour décrire une même entité.

Enfin, comme le modèle de données de data.bnf.fr se situe dans un environnement « ouvert » et « lié », il est également complété et enrichi par de nombreux alignements vers des référentiels extérieurs, tels que les lieux avec le référentiel Geonames[64], les personnes avec la base VIAF[65], ou encore les sujets Rameau avec le référentiel d'autorités de la Bibliothèque du Congrès[66].

[61] Voir http://xmlns.com/foaf/spec/
[62] Voir http://dublincore.org/documents/dces/
[63] Voir http://metadataregistry.org/schemaprop/list/schema_id/15.html
[64] Voir http://www.geonames.org/
[65] Voir http://viaf.org/
[66] Voir http://id.loc.gov/authorities/subjects.html

Après avoir expliqué comment était constitué le modèle de données RDF de data.bnf.fr, il convient à présent d'énumérer les constats et besoins liés au projet par rapport à la mise à jour et à la refonte de la documentation de celui-ci :

- **actualiser le modèle** : le projet data.bnf.fr étant piloté en méthode agile et ainsi en « version beta permanente »[67], il est nécessaire d'actualiser la documentation du modèle afin de la mettre en correspondance avec les évolutions fonctionnelles et développements informatiques de la dernière mise en production, dans un souci de cohérence et d'information des publics utilisateurs et surtout ré-utilisateurs de données du site.

- **améliorer la lisibilité du modèle** : le modèle s'enrichissant de nouvelles propriétés et d'alignements au fur et à mesure de son développement, il devient de plus en plus complexe à présenter et il apparait le besoin de mener une réflexion sur la présentation et la visualisation du modèle afin que des publics à la fois novices ou au contraire connaisseurs puissent le lire et le comprendre facilement suivant leurs besoins qui peuvent êtres variés (e.g. compréhension globale, une entité en particulier). Il convient alors de faire un effort particulier en direction des informaticiens développeurs, principaux réutilisateurs des données, qui ne sont pas familiers de certaines pratiques métiers. Ainsi doivent être mieux représentés visuellement le fonctionnement des FRBR, tout comme la distinction entre le « Concept » et l'« entité réelle » d'une même ressource doit être clairement explicité.

B.1.2 Modélisation des données RDF et SPARQL endpoint : deux outils complémentaires au cœur de la médiation avec les réutilisateurs

B.1.2.1 Documentation du modèle de données RDF, centré sur les réutilisateurs

Voici quelques exemples significatifs de mises à jour du modèle de données :

- Renseignement de la relation permettant de lier l'entité Œuvre à l'entité Expression avec la propriété « expressionOfWork » du vocabulaire rdarelationships,
- Evolution du nommage lexical du préfixe utilisé pour appeler le vocabulaire Dublin Core, de « dc » à « dcterms »,
- Documentation de la nouvelle entité « Evènement » implémentée dans data.bnf.fr, avec ses classes (de type « Work » du vocabulaire frbr-rda et « Event » du vocabulaire dcterms) et propriétés (e.g. pointage de l'entité Date par la relation exprimée avec la propriété « date » des vocabulaires dcterms et bnf-onto) associées,

[67] Voir http://en.wikipedia.org/wiki/Perpetual_beta

- Renseignement d'attributs supplémentaires de l'entité Œuvre, liés à l'intégration dans data.bnf.fr des périodiques (e.g. attributs « issn » du vocabulaire bibo et « frequency » du vocabulaire dcterms),
- Expression des nouveaux alignements internes (e.g. relation entre le Concept du Lieu et le Concept du Rameau Nom Géographique mis en correspondance, avec la propriété « closeMatch » du vocabulaire SKOS) et externes (e.g. relation entre le Concept du Lieu et son aligné dans les référentiels INSEE et IGN, avec la propriété « exactMatch » du vocabulaire SKOS).

La médiation a été également menée pour donner à voir le modèle de données RDF suivant différentes facettes et angles correspondant aux différents profils de publics que nous pouvons rencontrer dans data.bnf.fr, provenant du monde des bibliothèques comme d'autres communautés très variées[68]. Ainsi une vue d'ensemble, globale mais synthétisée, permet de voir en un coup d'œil l'ensemble des entités et classes du modèle ainsi que les principales relations et propriétés les liant entre elles (*figure* 1 de l'**annexe IV**).

La pédagogie de présentation du modèle se fonde ensuite sur le principe de *zooms* sur des aspects ou entités particuliers du modèle de données RDF de data.bnf.fr. Ces focus sont plus détaillés et exhaustifs et doivent permettre de répondre à des besoins et usages précis d'information et de données. Nous présentons par exemple exclusivement tous les alignements externes réalisés dans data.bnf.fr (*figure* 2 de l'**annexe IV**). Ce type d'information peut par exemple renseigner un décideur sur les « *réseaux* » avec lesquels la BnF s'est liée et associée[69], alors qu'une société de création d'applications web pourra voir tous les liens qui existent entre les données de la BnF et celles d'autres bases, ce qui lui permettra de construire des applications innovantes qui fédèrent ces données. Les autres zooms, permettent d'avoir une vue détaillée et complète des principales entités du modèle que sont le Concept, l'Œuvre, l'Auteur, le Lieu et la Date. Par exemple, des professionnels

[68] Parmi les réutilisateurs de data.bnf.fr figurent dans l'interprofession des bibliothèques, les bibliothèques municipales ou centres de documentation (ex : la bibliothèque de Fresnes inscrite dans projet OpenCat), des éditeurs SIGB qui visent à « FRBRiser » leur catalogue, intégrer RAMEAU à leur solution ou enrichir leur OPAC. Il y a également le monde de la recherche, avec l'exemple de la plateforme de recherche en sciences humaines et sociales Isidore ou Biblissima, observatoire du patrimoine écrit du Moyen Âge et de la Renaissance. De grandes institutions publiques sont également réutilisatrices de données comme le ministère des affaires étrangères qui avec IF Verso met en place une plateforme du livre traduit, ou encore Joconde Lab porté par le ministère de la culture. Il y a également les petites applications réalisées par des développeurs comme par exemple une frise chronologique des auteurs réalisée par Pierre Lindebaum ou le calculateur de domaine public de l'*Open Knowledge Foundation*. Parmi les réutilisateurs figurent également de grands acteurs du web tels que Wikipédia ou des sociétés administrant des moteurs de recherche.

[69] Illien, Gildas. « Décrire les objets du savoir, les nouveaux paradigmes du catalogage ».

travaillant sur des services web fondés sur les référentiels géographiques pourront directement s'intéresser au zoom sur le Lieu et y voir l'ensemble des propriétés, attributs et alignements implémentés dans data.bnf.fr (*figure* 6 de l'**annexe IV**).

La *figure* 3 de l'**annexe IV** permet de comprendre tout le sens que peut recouvrir l'entité Concept (Sujet Rameau, Auteur, Œuvre, Lieu, Evènement) et les attributs qu'elle contient (les données d'autorité catalographiques exprimées dans le vocabulaire SKOS que sont « prefLabel », « altLabel », « scopeNote », « editorialNote » et « note »).

La vue d'ensemble de la *figure* 1 est également une façon de représenter les différentes entités et relations du modèle FRBR global, alors que la *figure* 4, permet de bien visualiser et distinguer les relations et attributs des entités du FRBR permettant d'exprimer la production intellectuelle et artistique : Œuvre, Expression, Manifestation. Ces visualisations permettent de mieux comprendre le modèle conceptuel ainsi que de voir où et comment est renseignée l'information de manière logique.

B.1.2.2 Profilage de jeux de données pour le SPARQL endpoint de data.bnf.fr[70]

SPARQL, qui tire son nom de la comparaison métaphorique du terme anglais *sparkle* (l'étincelle)[71], est un standard du W3C pour le web sémantique. Il s'agit à la fois d'un protocole et d'un langage d'interrogation de données en RDF[72]. Il permet de réaliser des requêtes très fines et précises en exploitant toute la richesse des données structurées en RDF. Il permet également de faire des opérations algébriques comme des calculs arithmétiques ou différents tris sur les données. A ce jour, cet outil est principalement destiné à des informaticiens ou d'autres professionnels techniquement aguerris, car il suppose de savoir manier des opérateurs tels que SELECT, WHERE, COUNT, FILTER, REGEX, déjà connus dans le langage SQL d'interrogation de bases de données relationnelles, mais aussi de savoir manier des requêtes par appariement de graphes incluant une logique de raisonnement.

[70] Voir http://data.bnf.fr/sparql
[71] Voir http://fr.wikipedia.org/wiki/SPARQL
[72] Gandhon Fabien, Faron-Zucker Catherine, Corby Olivier. *Le web sémantique : Comment lier les données et les schémas sur le web ?*

Les requêtes SPARQL se font en exprimant des graphes représentés en triplets sous la forme Sujet – Prédicat – Objet, et intégrant des variables permettant des raisonnements par appariement de graphes. Prenons l'exemple de la requête suivante :

```
SELECT ?nom ?URI WHERE {
?URI rdf:type foaf:Person .
?URI foaf:gender "female" .
?URI rdagroup2elements:fieldOfActivityOfThePerson <http://dewey.info/class/100/> .
?URI foaf:name ?nom .
} ORDER BY ASC (?nom) LIMIT 3
```

Elle permet de rechercher dans le *triplestore* de data.bnf.fr des ressources correspondant à une Personne physique, du genre féminin, qui a pour domaine d'activité la philosophie. Ces informations sont indiquées dans le WHERE alors que les résultats que l'on souhaite voir nous retourner sont spécifiés dans le SELECT : il s'agit du nom et de son URI déréférençable. Enfin les dernières opérations permettent avec ORDER BY ASC d'ordonner les résultats des noms par ordre alphabétique croissant et avec LIMIT, de limiter le nombre de résultats à trois occurrences. En voici les résultats dans le format CSV :

```
"nom","URI"
"Angela Davis","http://data.bnf.fr/ark:/12148/cb11898827q#foaf:Person"
"Anneliese Maier","http://data.bnf.fr/ark:/12148/cb12359701f#foaf:Person"
"Annie Leclerc","http://data.bnf.fr/ark:/12148/cb11911948s#foaf:Person"
```

L'enjeu de réaliser des profilages de jeux de données (voir **annexe V**), a été de mettre en correspondance des besoins informationnels exprimés dans la langue naturelle avec leur correspondance en requête SPARQL. En voici un exemple :

Retrouver tous les titres du Roman de la Rose et la note associée :

```
SELECT DISTINCT ?titre_forme_internationale_francais ?formes_rejetees ?note_associee WHERE {
<http://data.bnf.fr/ark:/12148/cb166125510> skos:altLabel ?formes_rejetees ;
                                        skos:prefLabel ?titre_forme_internationale_francais ;
                                        skos:editorialNote ?note_associee }
```

Cette démarche permet d'accompagner les réutilisateurs dans leur entreprise d'exploitation du *triple store* de data.bnf.fr. Les personnes connaissant déjà data.bnf.fr et son modèle de données pourront s'appuyer sur les exemples de requêtes en SPARQL pour formuler à leur tour leurs propres requêtes, alors que ceux familiers des langages de requêtes seront plus attentifs aux requêtes exprimées en langage naturel et aux propriétés et attributs choisis pour récupérer l'information.

Le document permet ainsi de découvrir la base par les grands axes d'information qu'elle recèle, du point de vue des usagers : les informations sur une personne ou une organisation, les fonctions d'un auteur sur les documents (ex. : traducteurs, photographes, etc), l'indexation sujet Rameau, les identifiants (ARK[73], ISNI, ISBN, etc), les vocabulaires et les référentiels (genres musicaux, langue, pays, type de document, ouvrages jeunesse), les lieux et notices géographiques, les documents numérisés (liens vers Gallica), les illustrations d'une page data.bnf.fr.

B.2 Création de liens entre les ressources de la BnF, rapports humains / machines et évolution des pratiques de travail

B.2.1Les vocabulaires d'autorité au cœur de l'approche « Linked Entreprise Data »

Parmi les quatre premières étoiles qu'a tout d'abord proposées Tim Berners-Lee comme bonnes pratiques du *Linked Data*[74], figure le principe de « *linked RDF* », qui renvoie à la création de liens entre ressources[75]. Cette problématique, lorsqu'elle se pose *« entre jeux de données internes à une même organisation* » se nomme « *Linked Entreprise Data* » (LED)[76]. C'est dans cette optique que j'ai travaillé sur la fabrique de liens entre des données de la BnF cloisonnées : d'une part, des documents ayant un caractère géographique ayant été indexés selon les cas avec deux référentiels d'autorités distincts : autorités « Nom géographique » et autorités « Rameau Nom Géographique »). D'autre part les expositions virtuelles qui sont administrées dans une base de données isolée des autres bases de la BnF. Avec data.bnf.fr, le pivot documentaire entre les ressources, l'objectif est justement de faire du lien entre ces ressources, dans une démarche *Linked Entreprise Data*.

[73] Archival Resource Key
[74] Berners-Lee Tim. *Linked data*. URL : http://www.w3.org/DesignIssues/LinkedData.html
[75] "*Include links to other URIs. so that they can discover more things*"
[76] Bermès Emmanuelle, avec la coll. d'Isaac Antoine et Poupeau Gautier. Le Web sémantique en bibliothèque.

Comme nous le présente Emmanuelle Bermès, « *la clé du LED repose dans la capacité à réutiliser les données des différentes applications qui constituent le système d'information, tout en respectant les besoins métiers particuliers qui justifient l'existence de bases de données diverses* ». Il s'agit de construire ici une « *interopérabilité basée sur les liens* » et pour y parvenir, il est « *nécessaire d'identifier dans les données les entités qui jouent par nature un rôle de pivot entre les différentes bases* »[77]. Si les référentiels et vocabulaires d'autorités de la BnF jouent ce rôle de pivot entre les ressources, les deux cas d'application vus durant le stage ont l'intérêt de présenter des besoins et des logiques différentes.

B.2.2 Aligner automatiquement deux référentiels : travaux sur les autorités géographiques

Besoins : Rameau noms géographiques et autorités géographiques, des entités de la réalité identiques mais des « records » différents

Pour le cas des documents ayant un caractère géographique, il existe à la BnF deux référentiels d'indexation qui coexistent : d'une part, le langage documentaire RAMEAU (Répertoire d'autorité-matière encyclopédique et alphabétique unifié) qui contient la typologie des « noms géographiques » (NG), et d'autre part les autorités géographiques. Nous avons vus que les technologies du web sémantique nous permettaient de faire dialoguer des ressources hétérogènes grâce à un point de contact. Ces deux référentiels d'autorités constituent un cas d'école car ils portent sur le même type de données (géographiques) avec une structuration et une utilisation différentes. Voyons leurs caractéristiques :

RAMEAU[78], créée en 1980, est constitué d' « autorités matière » utilisées pour décrire le sujet contenu dans des documents stockés sur tout support, ainsi que pour gérer les points d'accès sujets aux notices bibliographiques. A la BnF, cela concerne les collections provenant du dépôt légal, des acquisitions ainsi que le catalogage rétrospectif ou la rétroconversion[79]. Le référentiel est constitué au fur et à mesure des besoins d'indexation, soit par les catalogueurs eux-mêmes (c'est notamment le cas des noms géographiques) soit par proposition faites par le réseau de ses utilisateurs (c'est le cas des noms communs). Il est donc élaboré et utilisé conjointement par la Bibliothèque nationale de France et d'autres établissements (bibliothèques universitaires, de lecture publique ou de recherche). Parmi ses types d'autorités gérées figurent les noms communs, les noms de personne, les collectivités, les titres de

[77] Bermès Emmanuelle, avec la coll. d'Isaac Antoine et Poupeau Gautier. Le Web sémantique en bibliothèque. p.159
[78] Voir http://guiderameau.bnf.fr/pdf/rameau_0007.pdf
[79] Voir http://fr.wikipedia.org/wiki/Catalogue_de_biblioth%C3%A8que

publication en série, les titres uniformes ou anonymes, les subdivisions chronologiques et les noms géographiques qui nous intéressent ici. Les noms géographiques Rameau concernent les noms de la géographie physique ou humaine, les mythes géographiques et les lieux imaginaires. Le schéma de métadonnées des autorités Rameau permet de renseigner la vedette (forme retenue), les termes exclus (formes rejetées), les liens sémantiques avec d'autres vedettes (termes associés, génériques ou spécifiques), les notes et les données de gestion. Les autorités Rameau sont exposées avec les technologies du web sémantique dans le format RDF/SKOS et sont accessibles dans data.bnf.fr sous la forme de pages HTML « thèmes » ainsi qu'en RDF/XML.

Le référentiel des autorités géographiques a été créé pour répondre aux besoins spécifiques d'indexation géographique des documents cartographiques. C'est en 1993 que la norme AFNOR NF Z 44-081, établie à partir des recommandations de la Commission de toponymie de l'Institut géographique national, définit les règles nationales pour l'établissement des vedettes noms géographiques, leur forme et leur structure. Son champ d'application concerne les documents cartographiques présents dans les collections de la BnF ainsi que tous les documents du Département des Cartes et plans nécessitant une entrée géographique. Au niveau de leurs métadonnées, les autorités géographiques ont une logique toponymique qui comble les manquements de Rameau sur ce point, en apportant plus de précision (e.g. coordonnées géographiques, divisions administratives) et de profondeur historique. Les notices d'autorité géographiques sont exposées avec les technologies du web sémantique avec la classe « spatialThing » de l'ontologie Geo[80] ainsi qu'en SKOS, et sont retrouvables dans data.bnf.fr à travers les pages HTML « Lieu » et ainsi qu'en RDF/XML.

Ces deux référentiels portent sur des objets conceptuellement proches, et il est donc pertinent de les aligner. Ils sont tous deux exposés selon les standards du web sémantique (URIs et RDF) qui permettent de les aligner dans une infrastructure « *linked data* ». A partir de ces éléments, il reste à réaliser pratiquement cet alignement. Face aux quantités (60 000 Rameau NG et 120 000 Géo), ainsi qu'à la qualité de normalisation et de structuration des données, une approche automatique de comparaison algorithmique de chaînes de caractères a été retenue.

[80] Voir http://smartreality.at/rdf/geo

Cependant, cette opération est rendue complexe en raison des différences dans la granularité des entités décrites, dans les schémas de métadonnées et normes utilisées, dans les pratiques de description et enfin dans les contraintes liées aux algorithmes eux-mêmes.

L'équipe projet data.bnf.fr et son prestataire avaient déjà réalisé un premier alignement Géo/Rameau NG en amont de mon arrivée en stage. Sur la page de lieu de data.bnf.fr « Paris »[81], nous retrouvons ainsi grâce à cet alignement tous les documents du département des cartes et plans (indexés avec des autorités géographiques) ainsi que tous les autres documents de la BnF (indexés avec des autorités Rameau NG). Cet alignement repose sur l'algorithme suivant :

- Dans un premier temps « îlotisation » par région puis par pays de sous-ensembles équivalents (ex. : « Ile-de-France » avec « Ile-de-France » ou « Etats-Unis » avec « Etats-Unis »)
- puis alignement au niveau du libellé des formes retenues ou rejetées (ex. : « Paris » avec « Paris »).

Cela permet de désambiguïser les nombreuses homonymies toponymiques telles que, Paris la capitale de la France[82] et Paris, la ville de l'Etat du Texas aux Etats-Unis[83]. Cet alignement a cependant des limites car il manque un certain nombre de rapprochements conceptuellement valides. En effet, cet algorithme ne prend pas en compte les règles métiers et pratiques de catalogage propres à ces deux référentiels, qui font que nous ne pouvons pas aligner tous les types de toponymes de cette manière (ex. : cours d'eau et massifs, quartiers, monuments, régions administratives, etc.). La mission a consisté à chercher des solutions pour aligner, parmi les 10 000 Géo non alignées avec des Rameau NG, celles qui devraient conceptuellement être alignées mais qui ne l'ont pas été. Et en conséquence, écrire des spécifications d'évolution de l'algorithme pour permettre de pallier ce problème de façon automatisée.

Solutions : de l'importance des données de contexte et de l'utilisation complémentaire des algorithmes et des ressources humaines

La méthodologie employée a consisté tout d'abord à réaliser un tableau de comparaison des schémas de métadonnées utilisés par les deux référentiels (voir **annexe VI**). L'objectif était d'identifier ce qui était comparable : pour pouvoir mettre en correspondance

[81] Voir http://data.bnf.fr/lieu/paris_france_/
[82] Voir http://data.bnf.fr/15282156/paris_france_/
[83] Voir http://data.bnf.fr/15240638/paris_texas_etats-unis_/

de manière pertinente et fiable, il est nécessaire de se fonder sur les propriétés discriminantes que vont contenir les entités des deux jeux de données[84]. Des entretiens avec les experts de ces deux référentiels m'ont permis de compléter cette approche en par les pratiques réelles des catalogueurs au-delà des normes (ex. : utilisation de telle zone ou non, cas particuliers ou pratiques courantes).

J'ai également intégré les transferts de compétences du prestataire sur les algorithmes d'alignements afin d'en comprendre les mécanismes : normalisation des chaînes de caractères (e.g. tokénisation, mots vides, caractères diacritiques…), *blocking*[85], choix de distances de comparaison (e.g. commence par, exact-match, Levenshtein[86], Jaccard[87]…), paramétrage (e.g. N-gram) et établissement de seuils. J'ai ainsi pu comprendre les atouts et les limites de ces algorithmes, et me familiariser par la pratique sur la manière d'utiliser à bon escient la panoplie de possibilités qu'ils offrent, en adéquation avec les référentiels et les jeux de données à aligner.

Ceci a conduit à l'écriture de spécifications pour l'amélioration de l'algorithme d'alignement Géo / Rameau NG (voir **annexe VII**). Tout d'abord, pour limiter les corpus à comparer et réduire le risque d'erreurs, une première étape de restriction du corpus aux seules données « alignables » a été proposée. Dans ce cas on ne cherche pas à rapprocher des données, mais plutôt à exclure des données n'ont aucune chance d'être alignées dans l'autre référentiel. Ainsi un test de type « contient un mot en commun » a été effectué pour chaque instance d'un référentiel avec l'ensemble des instances de l'autre référentiel. Réduire la taille des corpus permet enfin d'optimiser le temps de calcul des alignements, souvent coûteux en temps de calculs et en sollicitation de la machine.

Ensuite, il s'agissait d'améliorer la fiabilité des alignements en contextualisant les entités à rapprocher. Deux principaux types de données de contextes sont apparus : d'une part, tout ce qui se trouve à l'intérieur des villes[88] ; d'autre part, les territoires[89] et accidents géographiques[90]. Les deux référentiels renseignent ces informations dans des champs

[84] Isele, Robert et Bizer, Christian. « Active learning of expressive linkage rules using genetic programming ».
[85] Déf. : vise à restreindre les tentatives de comparaisons avec seulement les enregistrements pour lesquels un ou plusieurs identifiants discriminants sont présents et comparables.
[86] Voir http://fr.wikipedia.org/wiki/Distance_de_Levenshtein
[87] Voir http://fr.wikipedia.org/wiki/Indice_et_distance_de_Jaccard
[88] e.g. rue, quartier, grotte, place, temple, chapelle, etc.
[89] e.g. département, région, canton, empire, principauté, district, etc.
[90] e.g. cours d'eau, massif, estuaire, lac, courant, péninsule, etc.

discriminants et comparables. Par exemple les sous-zones du format Intermarc[91] « 170/470 $g » de Géo et « 167/467 $g » de Rameau NG, bien qu'elles correspondent à une désignation pour la première et à un qualificatif pour le second, permettent dans les deux cas de renseigner des informations de même nature. En utilisant ces données, nous allons reconstituer des lieux géographiques à aligner, en prenant l'hypothèse que deux instances peuvent être alignées si elles partagent deux éléments discriminant en communs. Soit par exemple un couple Géo[92] et Rameau NG[93] : ces deux notices partagent à la fois une caractéristique commune au niveau du libellé « Hierro » et du qualificatif : « île », elles ont donc une grande probabilité de correspondre au même objet de la réalité qui est l'« île de Hierro ».

Il faut également prendre en compte les pratiques d'indexation parfois différentes selon les cas et les langues. Ainsi toujours pour l'exemple de l' « île de Hierro », dans le référentiel Géo considérera que, pour la description de la forme en langue espagnole, la désignation en français ($g) est nécessaire : « spa $aIsla de Hierro$cCanaries$cEspagne$gîle » alors que dans la langue française cette donnée fait partie du nom du lieu et sera donc consignée dans la sous-zone correspondant à l'inversion du nom ($o) : « fre $aHierro$oÎle de$cCanaries$cEspagne ». Cette sous-zone de l'inversion du nom doit donc également être prise en compte dans le cas de l'alignement des territoires et accidents géographiques. Comme Rameau utilise la sous-zone $g (170 / 470) à la fois à des fins de localisation et de qualification, il s'est avéré nécessaire de créer une liste de termes destinée à nourrir l'algorithme afin de ne conserver que l'information utile, c'est-à-dire les territoires et accidents géographiques.

Les alignements Géo / Rameau s'avèrent enfin un cas d'école en ce qu'il montre une tension entre deux stratégies opposées : d'un côté, l'utilisation de règles très conservatrices qui ne feront que pas ou peu d'erreurs mais effectueront peu d'alignements ; de l'autre, le recours à des règles plus souples qui effectueront davantage d'alignements mais produiront alors plus de faux positifs. Cet enjeu pourrait se résoudre en ayant plus d'autonomie et de flexibilité sur la réalisation des alignements. Nous pourrions imaginer un processus par étapes successives, des plus conservatrices et sures, aux moins conservatrices, demandant une vérification humaine. La maîtrise d'un outil d'alignement comme Nazca[94] est une voie en ce sens.

[91] Voir http://www.bnf.fr/fr/professionnels/f_intermarc/s.format_intermarc_autorites.html?first_Art=non
[92] Voir http://catalogue.bnf.fr/ark:/12148/cb15292950k
[93] Voir http://catalogue.bnf.fr/ark:/12148/cb12134531v
[94] Voir http://www.logilab.org/project/nazca

Dans notre cas d'étude, nous pourrions par exemple faire un premier alignement très conservateur uniquement sur les formes retenues, puis un alignement moins conservateur prenant en compte et croisant les données de l'ensemble des formes rejetées des instances. Un opérateur verrait sa tâche facilitée et traiterait alors une liste réduite de candidats demandant validation, selon la méthode « *active learning* » décrite par Isel et Bizer[95].

B.2.3 Indexer manuellement des ressources ciblées : le cas des expositions virtuelles

Besoins : un manque d'accès normalisés et structurés, une approche qualitative nécessaire

Les expositions virtuelles de la BnF[96] sont des produits éditoriaux du service des Editions multimédias du département des Editions, au sein de la direction de la Diffusion culturelle. Ils se présentent sous la forme de sites thématiques et racontent une histoire tout en valorisant les collections papiers et surtout numériques de la bibliothèque. Ces sites et leurs composantes[97] s'appuient sur un schéma de métadonnées commun composé de 77 éléments, administrés dans un format Excel. Cependant il n'y a pas eu de catalogage systématique et normalisé pour indexer ces ressources avec des autorités BnF. Etant donné la richesse du contenu de ces ressources nécessitant un regard et une interprétation humaines pour évaluer et choisir une indexation pertinente, il a été choisi de réaliser manuellement les indexations manquantes avec des vocabulaires d'autorité.

Cette approche, à l'opposé du cas précédent, se situe dans la tradition professionnelle des bibliothèques où règne une culture forte de l'indexation manuelle. Elle représente un des piliers du métier ainsi que la valeur ajoutée du catalogue. C'est d'ailleurs sur cette activité que reposent historiquement la qualité et la fiabilité des données sources et la confiance que ces dernières suscitent auprès des publics et partenaires sur le web de données. Gildas Illien déclare à cet égard que, dans un environnement où l'interopérabilité globale des données, le *crowdsourcing* et le *datamining* appellent à repenser la pratique de catalogage, que celui-ci portera désormais sur un périmètre plus ciblé de ressources et que « *le traitement des documents portera moins sur leur description matérielle que sur l'analyse de leur*

[95] Isele, Robert et Bizer, Christian. « Active learning of expressive linkage rules using genetic programming ».
[96] Voir http://expositions.bnf.fr/
[97] Chaque exposition virtuelle suit la typologie suivante : un site est une ressource de type 1, qui contient des ressources de type 2 correspondant à des modules (e.g. album, conférences, chronologie, etc.) derrière lesquels on retrouve le cas échéant des ressources de type 3 correspondant à des notices grand formats (une unité d'information : une image en haute qualité avec son paratexte)

contenu »[98]. L'indexation des expositions virtuelles participe donc à cette idée de concentrer l'effort qualitatif sur certaines informations pour laquelle l'organisation (et le travail humain) présentent une forte valeur ajoutée par rapport à d'autres univers professionnels d'une part, et aux traitements automatisés d'autre part. Le recours aux vocabulaires d'autorité de la BnF est par ailleurs un gage d'interopérabilité, de contrôle et de pérennité de ces liens créés.

Solutions : de l'importe de l'analyse de contenu humaine et de l'interprétation

Le travail d'indexation manuelle des expositions virtuelles confirme l'importance cruciale de l'interprétation des contenus, à la fois pour identifier intelligiblement le sujet ou les sujets, mais aussi pour en trouver la correspondance la plus juste dans les vocabulaires d'autorité. Cela peut par exemple concerner les autorités personnes, lorsqu'il y a plus plusieurs homonymes dans le référentiel d'autorité et qu'il faut alors les désambiguïser avec les informations à disposition dans la notice d'autorité et dans l'exposition, voir aller chercher l'information ailleurs dans des sources sûres ; mais cela se démontre avec une acuité particulière avec l'indexation matière Rameau.

L'exemple suivant permet de l'illustrer : cinq ressources comprises dans le site de l'exposition Casanova[99] se succèdent avec les titres suivants : « Les cinq sens : le goût, l'esprit du vin »[100], « Les cinq sens : l'odorat, le parfum »[101], « Les cinq sens : le toucher, le corps dansé »[102], « Les cinq sens : la vue et le vêtement »[103], « Les cinq sens : l'ouïe, sens et sensualité de la voix »[104]. Nous interprétons et comprenons en analysant le titre de ces ressources et leur contenu, qu'il est moins question de « vin », de « parfum », de « corps dansé », de « vêtement » et de « voix » que plutôt des cinq sens de l'Homme que sont le goût[105], l'odorat[106], le toucher[107], la vision[108] et l'ouïe[109]. Nous comprenons aussi qu'il n'est pas question pour chacun des cas des « cinq sens » mais d'un seul précisément, tout comme nous avons bien saisi qu'il n'était pas question d' « esprit ». L'analyse du contenu par un humain et, au-delà, les choix d'indexation cohérents réalisés par un professionnel de l'information,

[98] Illien Gildas. « Décrire les objets du savoir, les nouveaux paradigmes du catalogage ». in Enjeux professionnels.
[99] Voir http://expositions.bnf.fr/casanova/
[100] Voir http://expositions.bnf.fr/casanova/conferences/01.htm
[101] Voir http://expositions.bnf.fr/casanova/conferences/02.htm
[102] Voir http://expositions.bnf.fr/casanova/conferences/03.htm
[103] Voir http://expositions.bnf.fr/casanova/conferences/04.htm
[104] Voir http://expositions.bnf.fr/casanova/conferences/05.htm
[105] Voir http://catalogue.bnf.fr/ark:/12148/cb11982973p
[106] Voir http://catalogue.bnf.fr/ark:/12148/cb11965395n
[107] Voir http://catalogue.bnf.fr/ark:/12148/cb11962402c
[108] Voir http://catalogue.bnf.fr/ark:/12148/cb11947019s
[109] Voir http://catalogue.bnf.fr/ark:/12148/cb12647533b

sont gages d'une qualité, d'un savoir-faire, d'une politique d'indexation et ainsi d'une plus-value qui généreront plus de pertinence dans l'accès aux ressources par une gestion raisonnée du bruit et du silence. On voit donc qu'un tel positionnement a du sens et de la pertinence s'il est appliqué de manière raisonnée, y compris dans le paysage évolutif des nouvelles pratiques « sœurs » du *crowdsourcing* et du *datamining*.

B.2.4 Affiner les algorithmes de « FRBRisation » des données : tests pour la réinjection dans le catalogue des liens calculés dans data.bnf

La culture du catalogage pratiqué manuellement[110] fait partie intégrante de l'identité professionnelle dans les métiers des bibliothèques et de « l'info-doc ». C'est ce savoir-faire et cette pertinence de l'indexation humaine qui apportent d'ailleurs une valeur ajoutée au catalogue et garantit la qualité, la fiabilité et la confiance en ses données ainsi que sa gestion performante du silence et du bruit. Pour préserver cette force et maintenir cette qualité de service dans un contexte où les effectifs en diminution obligent à repenser les modes de travail, la bibliothèque doit prendre en compte de manière l'arrivée des outils de traitement automatique des données en masse, afin de trouver un équilibre et une coopération adéquate et efficiente entre l'humain et la machine.

Dans ce contexte, le travail de réinjection dans le catalogue général de la BnF[111] des liens calculés par data.bnf.fr afin de « FRBRiser »[112] ce catalogue, a conduit le département de l'IBN à privilégier une approche où les experts humains évaluent les résultats des algorithmes de rapprochement, sur la base d'un panel de données représentatives, afin d'en améliorer le fonctionnement, plutôt que de tout corriger à la main. Il s'agit de développer une compétence d'ingénierie portant sur les algorithmes eux-mêmes dont les experts des données améliorent les spécifications en l'enrichissant de règles métiers.

Résumons tout d'abord le fonctionnement de l'algorithme utilisé dans data.bnf.fr pour réaliser les rapprochements entre publications et œuvres.

L'algorithme récupère tout d'abord tous les identifiants d'auteurs possédant à la fois des autorités titres et des titres de notices bibliographiques rattachées. A partir de cet ensemble et

[110] Illien Gildas. « Décrire les objets du savoir, les nouveaux paradigmes du catalogage ». in Enjeux professionnels.

[111] Le catalogue est l'outil de production de la BnF

[112] Il s'agit des liens entre des notices bibliographiques (des manifestations) et leurs autorités titres (l'œuvre)

pour chaque auteur, il va comparer les titres des notices d'autorité titres avec les titres des notices bibliographiques avec la comparaison « commence par »[113].

Ainsi une notice bibliographique correspondant à la manifestation du modèle FRBR (e.g. Harry Potter publié chez Gallimard en 2007[114]) va être raccrochée à une notice d'autorité titre correspondant au niveau œuvre du FRBR (Titre conventionnel Harry Potter[115]), comme le montre le schéma de l'**annexe VIII**.

Si cet algorithme fonctionne dans la plupart des cas généraux, nos tests nous ont permis d'identifier et de solutionner quelques problèmes dont nous présentons ici un cas signifiant : l'algorithme a rapproché ensemble des œuvres dérivées d'une même œuvre (e.g. le film « The Bells » de James Young[116] est aligné avec l'autorité titre de l'œuvre musicale « The Bells » de Kent Olofsson[117]). Le faux positif que constitue cet exemple s'explique par la présence d'un lien commun de ces deux notices à un même auteur : Edgar Allan Poe, auteur du texte du poème utilisé dans les deux autres oeuvres (les liens sont typés « texte de » ou « auteur adapté » selon les cas). Cette condition satisfaite, l'algorithme est passé à l'étape suivante de son *workflow* et a ainsi rapproché les chaînes de caractères de titres semblables (« The Bells »). Ce problème se répétant dans de nombreux cas similaires, il a été diagnostiqué que l'algorithme travaillait sur un corpus trop large produisant des erreurs. Pour améliorer la pertinence des résultats, il a été préconisé d'exclure les liens entre notices qui ramènent des œuvres adaptées.

Il est donc essentiel que les professionnels des bibliothèques s'approprient les mécanismes des algorithmes afin d'en améliorer la performance, en leur intégrant une intelligence métier indispensable pour résoudre les problèmes principaux rencontrés : faux positifs liés aux effets de bords, insuffisance de règles métier… Cette façon de voir les algorithmes parle de surcroît aux professions des sciences de l'information et des bibliothèques puisqu'elle pose les enjeux en termes de bruit et de silence documentaire.

[113] Les titres des notices bibliographiques étant parfois plus longs mais commencent la plupart du temps toujours par la même chaîne de caractère que l'œuvre, «Le Roman de la rose nouvellement revu et corrige » sera rapproché de « Le roman de la Rose ».
[114] Voir http://catalogue.bnf.fr/ark:/12148/cb410045886
[115] Voir http://catalogue.bnf.fr/ark:/12148/cb137554124
[116] Voir http://catalogue.bnf.fr/ark:/12148/cb414318200
[117] Voir http://catalogue.bnf.fr/ark:/12148/cb16530199n

C) Discussion du travail réalisé

Le travail réalisé a abouti à des apports concrets pour le développement du projet data.bnf.fr : la documentation du modèle de données RDF pour le focus sur le lieu est ainsi déjà en ligne sur la page web sémantique du site officiel[118], tout comme le document d'exemples de requêtes SPARQL sur lequel j'ai contribué[119]. Les spécifications d'amélioration de l'algorithme d'alignement Géo – Rameau (**annexe VII**) ont été transmises au prestataire informatique, sous la forme d'un bon de commande, et sont en cours de développement. Les indexations manuelles des expositions virtuelles feront elles l'objet d'une implémentation dans le site très prochainement.

Cette rapidité de production sur des aspects concrets du projet s'explique par les bénéfices apportés par la méthode agile, méthodologie employée par la BnF et son prestataire pour conduire le projet data.bnf.fr. En effet la méthode agile est une approche pragmatique qui permet une grande réactivité et adaptabilité des acteurs impliqués. Surtout elle permet d'avoir de nombreux échanges réguliers à différents intervalles, sous la forme de *scrums*[120] : c'est ainsi que j'ai pu échanger et avancer de manière quotidienne avec le soutien de mes tuteurs, la chef de projet de data.bnf.fr et le responsable des traitements automatisés. Au niveau hebdomadaire, nous avions des *scrums* organisés d'une part avec une équipe élargie d'experts du département de l'IBN et d'autre part avec l'équipe du DSI[121] en charge du projet data.bnf.fr et le prestataire informatique extérieur. Enfin les réunions de fin d'itération ainsi que les réunions du groupe de suivi métier élargies aux responsables de l'IBN, permettaient de faire des bilans et de se poser les questions stratégiques au niveau de l'établissement. L'ensemble de ces interactions aura été bénéfique et porteur.

Ce fonctionnement est exigent et demande un important investissement ainsi qu'une organisation de travail rigoureuse, afin de combiner travail personnel et réunions, action et prise de recul. Des points sur lesquels j'ai beaucoup appris mais qu'il sera nécessaire de consolider afin de continuer à progresser à l'avenir.

[118] Voir http://data.bnf.fr/docs/modele_lieux.jpg
[119] Voir http://data.bnf.fr/docs/doc_requetes_data.pdf
[120] Voir *en.wikipedia.org/wiki/**Scrum**_(software_development)*
[121] Département des Systèmes d'Information

D) Conclusion

La mission autour de la gouvernance des métadonnées de la BnF avec les technologies du web sémantique et du *record linkage* aura permis d'éclairer deux grands axes professionnels. D'une part, l'ouverture de la bibliothèque sur de nouveaux usages que sont les réutilisations des données ouvertes en RDF, provenant de communautés dépassant l'univers des bibliothèques. Cette démarche à conduit à renforcer la fonction de médiation entre les données exposées par la BnF et ses réutilisateurs, principalement à travers l'ouverture d'un SPARQL *endpoint*, mais aussi avec la production d'une documentation sur le modèle de données RDF de data.bnf.fr ainsi que sur des requêtes types et des jeux de données profilés.

D'autre part, les technologies du web sémantique et du *record linkage* ouvrent la voie à la création exponentielle de liens, en conformité avec les principes du *linked data* mais aussi du modèle conceptuel FRBR. Pour mener ces opérations, les professionnels de l'information voient leurs méthodes de travail évoluer. Ainsi l'affinage des algorithmes d'alignement de jeux de données avec des règles métiers devient une nouvelle façon d'aborder la description de ressources, répondant aux problématiques de traitements de masse et de contraintes économiques dans un contexte de réduction d'effectifs. Nous avons néanmoins pu mesurer que le travail d'indexation manuelle de ressources avec des vocabulaires d'autorité conservait, dans des cas ciblés et pertinents, toute sa valeur ajoutée, et restait donc indispensable pour maintenir la fiabilité et la confiance dans les professionnels des bibliothèques et dans l'institution.

A titre plus personnel, ce stage m'a permis de monter en compétence sur les technologies du web sémantique ainsi que sur l'alignement de jeux de données (*record linkage*). Il m'a aussi permis de mieux cerner les évolutions des pratiques des usagers comme des professionnels dans le contexte du web de données culturel. Enfin j'ai pu faire l'expérience concrète d'une méthodologie de travail en méthode agile, dont j'ai pu tirer un grand profit grâce à la qualité des équipes avec lesquelles j'ai travaillé à la Bibliothèque Nationale de France.

En conclusion, nous pouvons souligner deux sujets brûlants liés à ces missions : d'une part, l'intérêt croissant de la BnF pour ses réutilisateurs et réutilisations qui sont faites de data.bnf.fr, sujet récurrent pour tous les projets d'*open data*. D'autre part, les évolutions des pratiques professionnelles et la mise en place d'une coopération éclairée et efficiente entre l'homme et les algorithmes pour assurer une continuité de service de qualité, dans le contexte du web de données.

E) Bibliographie

Bermès Emmanuelle, avec la coll. d'**Isaac Antoine** et **Poupeau Gautier**. *Le Web sémantique en bibliothèque*. Paris : Editions du Cercle de la Librairie, 2013

Dalbin Sylvie. « Métadonnées et normalisation ». In *Métadonnées : mutations et perspectives. Séminaire INRIA 29 septembre-3 octobre 2008 – Dijon*. ADBS, 2008

Gandon Fabien, Faron-Zucker Catherine, Corby Olivier. *Le web sémantique : Comment lier les données et les schémas sur le web ?* Dunod, 2012

Groupe d'incubation W3C. *Rapport final du groupe d'incubation « Bibliothèques et Web de données » - Rapport du 25 octobre 2011*. trad. (sept. 2012) de « Library Linked Data Incubator Group Final Report ». Consulté le 01/08/2014, [en ligne] : http://mediatheque.cite-musique.fr/MediaComposite/ARTICLES/W3C/XGR-lld-fr.html

Groupe de travail IFLA. Fonctionnalités requises des notices bibliographiques : rapport final. http://www.bnf.fr/documents/frbr_rapport_final.pdf

Illien Gildas. « Décrire les objets du savoir, les nouveaux paradigmes du catalogage ». in *Enjeux professionnels.* Documentaliste-Sciences de l'information. Vol. 50, 2013

Isele, Robert et **Bizer, Christian**. « Active learning of expressive linkage rules using genetic programming ». in *Web Semantics : Science, Services and Agents on the World Wide Web*. Elsevier, 2013

Poupeau Gautier. « RDF,la grammaire universelle des machines ». ». in *Enjeux et technologies : des données au sens*. Documentaliste-Sciences de l'information. Vol. 45, 2011.

Vatant Bernard. « Des métadonnées à la description de ressources : les langages du Web sémantique ». In *Métadonnées : mutations et perspectives. Séminaire INRIA 29 septembre-3 octobre 2008 – Dijon*. ADBS, 2008

Wenz Romain. « Data.bnf.fr : au-delà des silos ». *Approches documentaires : priorité aux contenus*. Documentaliste-Sciences de I'Information vol. 48, n 4 ; 2011

F) Sommaire des annexes

Annexe I : modèle FRBR
(Source : http://www.frbr.org/2006/02)

Annexe II : Mapping formats originaux vers vocabulaires RDF
- Aurorités titres (MARC / RDF)
- Titre archives et manuscrits (EAD / RDF)

Annexe III : De la notice à la donnée

Annexe IV : Documentation du modèle de données RDF
- Figure 1 : vue d'ensemble
- Figure 2 : alignements vers des jeux de données externes
- Figure 3 : focus sur le concept
- Figure 4 : focus sur l'œuvre
- Figure 5 : focus sur l'auteur
- Figure 6 : focus sur le lieu
- Figure 7 : focus sur la date

Annexe V : Profilage de jeux de données : quelques requêtes types avec SPARQL *endpoint* pour interroger le *triplestore* de data.bnf.fr
(Source : http://data.bnf.fr/docs/doc_requetes_data.pdf)

Annexe VI : tableau de comparaison des schémas de métadonnées employées pour les référentiels Rameau Noms Géographiques et Géo

Annexe VII : spécifications pour l'amélioration de l'algorithme d'alignement Géo / Rameau NG

Annexe VIII : alignement des notices bibliographiques avec des notices d'autorité titres afin de « FRBRiser » le catalogue

Dec., 1946 :

« Each person in the world creates a Book of Life.
This book starts with birth and ends with death.
Its pages are made up of the records of the principal events in life.
Record linkage is the name given to the process of
assembling the pages of this Book into a volume. »

Halbert L. Dunn
Chief, National Office of Vital Statistics,
U. S. Public Health Service, Federal Security Agency,
Washington

Partie II. Rapport de recherche : Les alignements de vocabulaires d'autorité sur le web de données culturel, à la croisée d'enjeux technologiques, professionnels et stratégiques

A) Introduction

Le terrain de stage nous a permis de mettre en pratique une nouvelle gouvernance des données catalographiques de la BnF sur le web de données. Nous avons vus que nous nous placions alors dans un écosystème ouvert et lié où les données sont l'unité documentaire atomique. Lier ses données à d'autres étant le principal fondamental du *linked data,* comme c'est le cas par exemple dans data.bnf.fr avec des alignements avec Geonames, Wikipedia, la Library of Congress, etc. Cette pratique se répandant à travers de nombreux projets et remettant en question profondément bon nombre d'habitudes et de canevas, il est apparu nécessaire d'effectuer un travail de fond afin de problématiser le sujet des alignements de vocabulaires d'autorité sur le web de données culturel. Ceci afin de découvrir et de comprendre, à travers et au-delà du projet data.bnf.fr, quels sont les enjeux et les questions que soulèvent ce sujet du point de vue des nombreux acteurs, projets et positionnements différents. Le sujet des alignements des vocabulaires d'autorité sur le web de données culturel a été particulièrement intéressant à étudier en travail de recherche pour le caractère stratégique qu'il recèle et sa propension à faire cohabiter et interagir ensemble des questions d'ordre techniques, politiques, économiques, professionnelles, et de pratiques et d'usages.

B) Les alignements de vocabulaires d'autorité sur le Web de données culturel : des enjeux technologiques, professionnels et stratégiques

Définition de la problématique

Les alignements de vocabulaires d'autorité sur le Web de données, par les institutions culturelles, sont au carrefour d'enjeux multiples, à la fois technologiques, professionnels et stratégiques, ce qui nécessite de bien en cerner la problématique.

Tout d'abord, il convient de bien spécifier ce que l'on entend par vocabulaires d'autorité, tout en précisant ce qui ne rentre pas dans le champ de notre sujet de recherche (e.g. l'alignement d'ontologies). Ensuite nous expliquerons pourquoi l'alignement de vocabulaires d'autorité est un levier pour rendre interopérables des bases de données hétérogènes sur le Web de données. Ceci posé, nous nous intéresserons également à la façon dont ces alignements sont réalisés. Ces éléments permettent de poser le cadre technologique du sujet. Nous devons également nous pencher dans un second temps sur l'exploitation qui peut être faite de ces alignements, tant dans une perspective de production que de diffusion des données ainsi que de la réception par les utilisateurs. Enfin il convient d'étudier les enjeux stratégiques liés à la création et au maintien de ces alignements. On abordera ainsi les notions de confiance, de valeur et de

provenance de l'information sur le Web de données, ainsi que les problématiques de gouvernance des alignements. Cet aspect conduit à des interrogations de nature aussi bien technique que politique.

Notre problématique se place donc au carrefour de multiples domaines, dans un contexte technologique, professionnel et stratégique en pleine mutation : technologique, avec le recours aux standards du Web sémantique ainsi qu'aux algorithmes de rapprochement de données ; professionnel, avec la question de la production et diffusion des données comme celle de la réception par les usagers ; stratégique, où acteurs et décideurs doivent mettre en place une nouvelle gouvernance avec de nouveaux outils qui en redéfinissent les possibilités.

Positionnement de la problématique dans la littérature scientifique et professionnelle

Nous allons à présent explorer plus en détail le positionnement dans la littérature scientifique et professionnelle de la problématique des alignements des vocabulaires d'autorité sur le Web de données culturel.

1. Aligner les vocabulaires d'autorité des institutions culturelles : les raisons et la problématique de la réalisation

1.1. De la contrainte de l'existant à l'opportunité : valeur des vocabulaires d'autorité des institutions du patrimoine culturel à l'heure du web de données

Situons d'abord l'objet de ce que nous désignons lorsque nous parlons d'alignements de vocabulaires d'autorité sur le Web de données culturel. Il faut d'abord replacer dans une perspective historique la question des alignements entre vocabulaires d'autorité sur le web de données culturel. Le web de données ne fait par table rase du passé et il est essentiel de le penser en continuité avec ce qui l'a précédé, c'est ce que Sylvie Dalbin nomme la «*contrainte de l'existant* »[122] : lorsqu'il s'agit d'appliquer de nouveaux systèmes d'information développés dans le monde de la recherche à leurs environnements d'exploitation, le maintien de la continuité de service est une obligation. L'auteur rappelle aussi qu'au-delà de la contrainte, c'est également dans cet existant que se trouve « *la matière première et le moteur des activités* »[123] dès lors que l'information est au cœur de l'activité de l'organisation concernée. C'est dans cet esprit que les bibliothèques ont vu dans certaines de leurs données non pas une contrainte, mais une opportunité nouvelle d'exister et de jouer un rôle prépondérant au sein de l'environnement jeune et porteur du Web de données. Le rapport du groupe d'incubation du

[122]Dalbin Sylvie. « Représentation et accès à l'information : transformation à l'œuvre ». p.23
[123]Ibidem.

W3C « Bibliothèques et web de données » décrit ces données de bibliothèques de la manière suivante : « *données bibliographiques, autorités, vocabulaires conceptuels* »[124]. Plus précisément, les données qui nous intéressent ici dans le cadre de notre recherche sont ce que le rapport du W3C nomme les « *vocabulaires d'autorité* »[125]. Nous pouvons qualifier ces données, en reprenant la formule de Katell Briatte, de « *chair* »[126] du système d'information des bibliothèques. Françoise Bourdon et Vincent Boulet rappellent que ces données d'autorité, qui sont contrôlées et sourcées, ont été « *établies par les bibliothèques pour gérer les points d'accès aux notices décrivant les documents recensés dans leurs catalogues* ». Aujourd'hui « *libérées* »[127] de leurs notices et exposées sur le Web de données, elles jouissent d'une « *réputation de fiabilité et sont considérées dans le cadre du Web sémantique comme étant des données de confiance* »[128], dans un écosystème ouvert et lié où par exemple « *les données d'identification d'une personne, d'une collectivité ou d'une œuvre prennent une importance jusqu'alors inégalée dans le passé* »[129].

Pour être précis, complet et clair, dans un domaine où la terminologie est encore fluctuante, notre travail porte sur ce qu'Antoine Isaac décrit comme un type de référentiel renvoyant aussi bien aux « *vocabulaires de valeurs* » qu'aux « *Systèmes d'Organisation des Connaissances* »[130]. Il s'agit des listes d'autorités (ex. : noms de personnes, de collectivités, titres d'œuvres) et parmi les SOC, principalement des « *langages documentaires et thésaurus* » et « *des classifications épistémiques universelles de la bibliothéconomie et les approches à facettes universelles* »[131]. Il ne s'agit par contre pas des ontologies qui, bien qu'également classées parmi les SOC par Manuel Zacklad[132], relèvent d'une autre logique (classes, propriétés et relations entre elles) et d'autres objectifs qui font qu'elles « *n'ont pas une vocation exclusivement documentaire au sens de l'indexation et de la recherche d'information mais elles visent aussi à participer à l'ingénierie des connaissances d'un domaine* »[133]. Il ne s'agit pas non plus des « *éléments de description des métadonnées* »[134] ou schémas de métadonnées qui ne relèvent pas des « *données de valeurs* » mais plus de

[124]Groupe d'incubation W3C. Rapport final du groupe d'incubation « Bibliothèques et web de données ».
[125]Ibidem.
[126]Briatte Katell. « Hadoc, un programme pour harmoniser les données culturelles ». In Documentaliste - Sciences de l'information
[127]Bourdon Françoise, Boulet Vincent. VIAF : un pivot pour l'accès multilingue à diverses collections.
[128]Ibidem.
[129]Ibidem.
[130]Isaac Antoine. « Les référentiels : typologie et interopérabilité ».
[131]Zacklad Manuel. Évaluation des systèmes d'organisation des connaissances.
[132]Ibidem.
[133]Ibidem.
[134] Bermès Emmanuelle, avec la coll. d'Isaac Antoine et Poupeau Gautier. Le Web sémantique en bibliothèque.

l'organisation et de la présentation de celles-ci. Ces vocabulaires font également l'objet d'alignements sur le web de données, mais relèvent d'autres enjeux, pour reprendre toujours la formulation de Katell Briatte, davantage du côté de l'« *ossature* » que de la « *chair* »[135], ce qui correspond à un autre niveau d'interopérabilité, qui touche beaucoup plus les systèmes d'information.

Les vocabulaires d'autorité des institutions du patrimoine culturel, sont particulièrement intéressants sur le Web de données lorsqu'il s'agit de « *construire l'interopérabilité entre des données issues de domaines différents* »[136]. Cette interopérabilité basée sur les liens, permet ainsi de créer des points de contact entre des bases et jeux de données hétérogènes, sans chercher à avoir des modèles conceptuels, des schémas de métadonnées ou des ontologies en commun. Cela permet, comme le montre Emmanuelle Bermès, de lier les données sans perdre les « *particularités métiers* »[137] qui peuvent exister par exemple entre les approches provenant des musées, des archives, ou des bibliothèques dans la culture ou encore dans d'autres domaines. Cela entre bien en résonance avec les préconisations du groupe d'incubation du W3C, qui vise à rendre les données des institutions patrimoniales culturelles, notamment les vocabulaires d'autorité (listes d'autorités ainsi que vocabulaires conceptuels), plus visibles et réutilisables, dans un contexte plus large que celui d'origine[138].

1.2. La réalisation des alignements de vocabulaires d'autorité : différentes stratégies oscillant entre le manuel et l'automatique

Comme nous l'indique Antoine Isaac, les alignements de vocabulaires d'autorité sont plus difficiles qu'entre ontologies car ils sont à la fois plus volumineux et moins structurés[139]. Le « *multilinguisme* »[140], comme comme « *les choix de catalogage locaux* »[141] pour la construction de ces vocabulaires d'autorité sont en effet des enjeux majeurs dont il faut tenir compte, lorsque l'on souhaite pratiquer un alignement. Ainsi, selon les contraintes, différentes méthodologies d'alignement sont possibles. Si dans le cas du programme MACS[142] (Multilingual ACcess to Subjects : Accès multilingue par sujet) le choix est toujours fait sur

[135]Briatte Katell. « Hadoc, un programme pour harmoniser les données culturelles ». In Documentaliste - Sciences de l'information

[136]Bermès Emmanuelle et Poupeau Gautier. "Les technologies du web appliquées aux données structurées".

[137] Bermès Emmanuelle, avec la coll. d'Isaac Antoine et Poupeau Gautier. Le Web sémantique en bibliothèque.

[138]Groupe d'incubation W3C. Rapport final du groupe d'incubation « Bibliothèques et web de données ».

[139]Isaac Antoine. Les référentiels : typologie et interopérabilité.

[140]Idem.

[141]Bourdon Françoise, Boulet Vincent. VIAF : un pivot pour l'accès multilingue à diverses collections.

[142]Le programme MACS a pour but de développer l'accès multilingue par sujet, afin de permettre à l'utilisateur final ou professionnel d'interroger directement, dans sa langue maternelle, le contenu de catalogues étrangers, quelque soit la langue d'indexation.

un alignement manuel, la publication du thésaurus multilingue AGROVOC[143] par la FAO (Food and Agriculture Organization) des Nations Unies a conduit au choix de l'alignement semi-automatique, afin « *de combiner l'utilisation de systèmes de mise en correspondance automatique avec une évaluation manuelle intensive des résultats par des experts de domaine* ».[144] En effet, les outils d'alignements automatiques ne fournissent pas toujours des résultats complètement satisfaisants et de nombreux outils d'alignement semi-automatiques ont été développés[145].

Isele et Bizer nous exposent l'intérêt d'une stratégie de type « *active learning* »[146], en nous montrant qu'elle facilite le travail du producteur de liens, en lui proposant une liste réduite de candidats à l'alignement qui demanderont une validation. Pour comprendre ces enjeux, il faut aussi prendre en compte le fonctionnement des outils de mise en correspondance automatique de chaînes de caractères pour en repérer les avantages et les limites.

Si nous reprenons la méthodologie qui nous est présentée par Isele et Bizer, nous pouvons noter que les systèmes de mise en correspondance se fondent sur les propriétés discriminantes que vont contenir les entités d'un jeu de données. Ensuite ces systèmes, après avoir transformé les chaînes de caractères pour les rendre comparables (ex.: normalisation, tokénisation, concaténation), utilisent des mesures de distance ainsi que des seuils de distance appropriés pour effectuer leurs calculs.

Ces mesures de distances peuvent s'appliquer au niveau des caractères (ex : distance de Levenshtein), aux mots (ex : coefficient de la distance de Jaccard), aux nombres (ex : la différence numérique), à la géographie (ex : la distance géographique en kilomètres) ou aux dates (ex : la différence entre deux dates)[147]. Ces méthodologies fondées sur le calcul sont ainsi très utiles à l'homme. Elles doivent néanmoins être utilisées à bon escient, pour effectuer des traitements en masse industriels tout en permettant la réalisation de services de bonne qualité.

[143]Voir http://aims.fao.org/fr/standards/agrovoc
[144]Isaac Antoine. Les référentiels : typologie et interopérabilité.
[145]Isele, Robert et Bizer, Christian. « Active learning of expressive linkage rules using genetic programming ».
[146]Ibidem.
[147]Ibidem.

Une fois les rapprochements réalisés, l'ontologie SKOS[148] permettra de représenter dans le formalisme RDF le « *lien d'équivalence conceptuelle exacte* » avec la propriété « exactMatch » et le « *lien d'équivalence conceptuelle approximative* » avec la propriété « closeMatch ». Enfin, la propriété « sameAs » de l'ontologie OWL est une propriété qui permet d'exprimer la coréférence[149].

2. L'exploitation des alignements du point de vue des professionnels comme du point de vue des utilisateurs

2.1. Perspectives pour la production des données

2.1.1 Mutualiser l'effort de catalogage et rationaliser les coûts

Aligner des vocabulaires d'autorité dans le secteur culturel peut, comme le déclare Katell Briatte, conduire à construire un écosystème inter-institutionnel où l'effort de production des métadonnées descriptives est mutualisé, « *tant en termes de charge de travail que de complétude des données* »[150]. Gildas Illien nous indique que les institutions telles que les bibliothèques pourront dans cet environnement, « *concentrer leur production bibliographique interne sur les ressources qui constituent leur valeur ajoutée propre* »[151]. Ces perspectives s'inscrivent en continuité avec les « *modèles de partage collaboratifs utilisés de longue date par les bibliothèques* »[152]. Ainsi est évoquée, par Emmanuelle Bermès, l'idée d'un « *catalogage partagé »* réinventé[153].

Les alignements de vocabulaires contrôlés, conduisant à mutualiser l'effort de catalogage, rentrent également dans une logique économique de « *rationalisation des coûts* »[154]. Ils représentent d'ailleurs de ce point de vue là aussi bien des avantages, que simplement une « *nécessité* »[155] dans le contexte de la Révision Générale des Politiques Publiques (RGPP) engagée depuis 2007.

[148]Gandon Fabien, Faron-zucker Catherine, Corby Olivier. Le Web sémantique : Comment lier les données et les schémas sur le web ?
[149]Vatant Bernard. « Des métadonnées à la description de ressources : les langages du web sémantique »
[150]Briatte Katell. « Hadoc, un programme pour harmoniser les données culturelles ».
[151]Illien Gildas. « Décrire les objets du savoir, les nouveaux paradigmes du catalogage ».
[152]Groupe d'incubation W3C. Rapport final du groupe d'incubation « Bibliothèques et web de données ».
[153] Bermès Emmanuelle, avec la coll. d'Isaac Antoine et Poupeau Gautier. Le Web sémantique en bibliothèque.
[154]Ibidem.
[155]Illien Gildas. « Décrire les objets du savoir, les nouveaux paradigmes du catalogage ».

2.1.2 Consolider collaborativement la fiabilité de l'identification des autorités

Un autre apport essentiel de ces alignements de vocabulaires d'autorité est au niveau de la consolidation de ces données d'autorité. C'est à dire, comme l'évoque Katell Briatte, il devient par exemple « *possible d'envisager la consolidation des données d'identification des biens culturels dans une carte d'identité unique* ». Les « *données de référence* » ainsi constituées offrent alors des garanties d' « *authenticité* », « *d'unicité* », de « *qualité* », « *d'exhaustivité* » et de « *pérennité* » aux producteurs comme aux usagers. Cela conduira, dans le cadre de la brique d'« harmonisation des vocabulaires » du programme Hadoc[156], à la création d'un « *référentiel des biens culturels* ».[157]

Ce type de modèle permettant de consolider les vocabulaires d'autorité en alignant les référentiels en un endroit centralisé, comme c'est le cas par exemple avec le projet VIAF[158], correspond au modèle « *hub and spoke* » décrit par Emmanuelle Bermes[159]. Il faut noter également que l'application GINCO[160], fournissant un environnement pour la gestion des vocabulaires du Ministère de la Culture et de la Communication, s'appuie sur la norme ISO 25964-1:2011[161] (Thésaurus et interopérabilité avec d'autres vocabulaires) qui formalise « *la distinction* [...] *entre d'une part les concepts et d'autre part les termes les représentant* »[162] . Au-delà de l'alignement, il devient alors possible de réaliser des *clusters* permettant de regrouper l'ensemble des termes alignés derrière un unique concept et donc, dans le web sémantique, une seule URI.

Dans le cadre du projet VIAF[163], ce sont les vieux principes du Contrôle Bibliographique Universel[164] qui sont revisités grâce aux technologies du Web sémantique et au paradigme de l'open data[165]. Chaque partenaire, agence bibliographique nationale mais aussi en dehors du monde des bibliothèques : musée (Getty Institute[166]), Wikipédia, et bientôt d'autres acteurs du monde de la documentation, fournit ses données d'autorité sur des personnes, des organisations, des œuvres, des expressions, des *meetings*, des lieux.

---

[156]Voir http://www.culturecommunication.gouv.fr/Ressources/HADOC/Referentiels2/Les-vocabulaires-scientifiques-et-techniques/L-harmonisation-des-vocabulaires

[157]Briatte Katell. « Hadoc, un programme pour harmoniser les données culturelles ».

[158]Voir www.oclc.org/viaf/

[159] Bermès Emmanuelle, avec la coll. d'Isaac Antoine et Poupeau Gautier. Le Web sémantique en bibliothèque.

[160]Gestion Informatisée de Nomenclatures Collaboratives et Ouvertes

[161]Briatte Katell. « Hadoc, un programme pour harmoniser les données culturelles ».

[162]Dalbin Sylvie, Yakovleff Nathalie, Zysman Hélène. et. al. Livre blanc : ISO 25964-1 – Thésaurus pour la recherche documentaire

[163]Virtual International Authority File

[164]Concept de partage de la production de données en bibliothèque à l'échelle mondiale grâce à l'apport de chaque bibliographie nationale

[165]Angjeli Anila, Mac Ewan Andrew, Boulet Vincent. ISNI and VIAF – Transforming ways of trustfully consolidating identities.

[166]Voir http://www.getty.edu/

Ces données d'autorités font alors l'objet d'agrégations après avoir été rapprochées algorithmiquement, et permettent ainsi d'identifier les autorités avec confiance. Dans cet écosystème ouvert et lié, les données d'autorité sont donc consolidées et enrichies par chaque partenaire de VIAF. Comme le décrivent Anila Angjeli, Andrew Mac Ewan et Vincent Boulet, une nouvelle ère de consolidation collaborative des identités à l'échelle mondiale s'ouvre[167].

2.2. Perspectives pour la diffusion des données

2.2.1 « Rencontrer » les pratiques des internautes sur le Web et favoriser la navigation et la sérendipité

Les pratiques des publics sur le Web débouchent sur le constat que la logique institutionnelle n'est plus prépondérante, les internautes souhaitant accéder à de l'information de « *manière unifiée et indifférente aux logiques disciplinaires ou organisationnelles qui ont présidé leur création* »[168]. En effet, comme cela nous est expliqué par Emmanuelle Bermès, « *sur le Web, la démarche de l'internaute n'est pas centrée sur les institutions mais sur les contenus* »[169]. Cet aspect des pratiques rend ainsi essentielle la convergence des institutions culturelles[170], ce que les alignements entre vocabulaires d'autorité nous l'avons vu nous permettent de faire.

Pierre Col parle d'un « *mode de consommation de la culture et de son patrimoine* » où l'internaute « *ne part pas d'une action délibérée qui implique la visite d'un site précis, mais d'opportunités en rapport avec des recherches personnelles ou liées à des études ou en relation avec un intérêt ponctuel ou régulier* »[171]. Ce constat débouche sur l'idée de « *provoquer la sérendipité* »[172], chose qui est facilitée par la création de liens entres les ressources de diverses institutions.

Également, ces nouvelles pratiques s'inscrivent dans une autre, très répandue, qui est que « *l'internaute privilégie certaines grandes portes d'entrée sur le Web que sont les moteurs de recherche (...), les réseaux sociaux (...) ou encore Wikipédia* »[173]. Ce constat conduit à renforcer l'intérêt et même la nécessité pour les institutions culturelles de se lier à d'autres partenaires, plus éloignés de leur inter profession afin de se placer sur le «*chemin naturel des*

[167]Angjeli Anila, Mac Ewan Andrew, Boulet Vincent. ISNI and VIAF – Transforming ways of trustfully consolidating identities.
[168]Briatte Katell. « Hadoc, un programme pour harmoniser les données culturelles ».
[169]Bermès Emmanuelle. « Bibliothèques, archives et musées : l'enjeu de la convergence des données du patrimoine culturel. »
[170]Ibidem.
[171]Col Pierre. Culture et patrimoine sont à la pointe du Web sémantique.
[172]Ibidem.
[173]Ibidem.

internautes »[174]. C'est notamment le sens que peut recouvrir les alignements réalisés par les institutions culturelles avec les projets Wikipédia et le jeu de données correspondant, DbPedia.

Les alignements de vocabulaires d'autorité, en ce sens, facilitent l'accès à diverses collections ou ressources indépendamment de leur nature ou de leur production d'origine. Une vertu notable de cet écosystème de liens est qu'il peut permettre de mettre en valeur des sources jusque-là peu ou inconnues[175], grâce à l'alignement de sources à l'audience disparates.

Pour conclure ce point, reprenons Lorcan Dempsey qui déclare qu'il est important pour les bibliothèques de se placer dans des réseaux décentrés d'eux-mêmes, afin d'être présent sur le flux informationnel et communicationnel de ses usagers (« *to be in the flow of it users* »). Il nomme ceci la tendance centrifuge, en référence au mécanisme d'éloignement de son centre (e.g. sur les réseaux sociaux ou d'autres environnements via l'agrégation et la syndication des données, contenus et services de la bibliothèque), et en opposition à la tendance centripète, orientée sur une présence exclusivement institutionnelle (e.g. le site web de la bibliothèque)[176].

2.2.2 Faciliter la recherche d'information et la découverte par l'enrichissement sémantique et le multilinguisme

Les liens entre les données d'autorité vont également augmenter la portée des index et rendre les recherches fédérées plus performantes, ce qui aura pour conséquence d'offrir aux utilisateurs de nouveaux moyens de « *découverte* » ainsi que « *des possibilités de navigation plus riche* »[177]. Ces alignements de vocabulaires d'autorité resteront complètement invisibles en tant que tels pour l'utilisateur[178] mais lui apporteront de la valeur ajoutée dans ses recherches et sa navigation.

Les alignements des données d'autorité et en particulier des SOC vont permettre, comme l'indique Jean Delahousse, l'« *alimentation des lexiques des moteurs de recherche* »[179]. En effet ces derniers utilisant des lexiques pour l'indexation des ressources documentaires, l'alignement de deux SOC va permettre d'enrichir ces lexiques et ainsi « *d'accéder à l'information avec un vocabulaire plus riche et couvrant les deux référentiels* ». Par exemple,

[174] Bermès Emmanuelle, avec la coll. d'Isaac Antoine et Poupeau Gautier. Le Web sémantique en bibliothèque.
[175] Groupe d'incubation W3C. Rapport final du groupe d'incubation « Bibliothèques et web de données ».
[176] Lorcan Dempsey's Weblog. The decentred library network presence. 20 avril 2014. http://orweblog.oclc.org/archives/002216.html
[177] Ibidem.
[178] Bermès Emmanuelle, avec la coll. d'Isaac Antoine et Poupeau Gautier. Le Web sémantique en bibliothèque.
[179] Delahousse Jean. « Sur l'alignement et la mise en correspondance de terminologies »

en fournissant à un moteur de recherche un lexique décrivant les concepts à la fois avec les thésauri Garnier (photos) et Rameau (livres), un utilisateur pourra utiliser indifféremment les termes de l'un ou de l'autre référentiel pour accéder au même concept[180]. L'utilisateur pourra aussi profiter des extensions sémantiques offertes par l'alignement de thesauri différents, en voyant sa recherche élargie ou des suggestions connexes plus variées : plus de termes synonymes, spécifiques ou génériques. Notons que les alignements peuvent également s'opérer directement au niveau du texte intégral. Ainsi par exemple, le projet ISIDORE[181] utilise de nombreux référentiels (le référentiel HAL, l'index des catégories thématiques d'Open Edition, le vocabulaire Rameau, les thésaurus Pactols, GEMET, GéoEthno et enfin le référentiel géographique Geonames) pour enrichir ses données en comparant par des algorithmes de rapprochement les métadonnées ou le texte intégral des ressources avec les entrées des différents référentiels.

Dans une perspective de multilinguisme, l'alignement de vocabulaires d'autorité peut également s'effectuer sur des listes d'autorités ou des terminologies traitant d'un même domaine dans des langues différentes. Le projet VIAF permet par exemple de relier des données d'autorités multilingues, provenant de nombreuses bibliothèques nationales et grandes institutions culturelles, entre elles et joue la fonction de « *tremplin entre données* » en étant « *utilisable comme rebond vers d'autres ressources* ». Ainsi Françoise Bourdon et Vincent Boulet notent que VIAF permet « *l'interopérabilité entre des données conçues selon des règles différentes, dans des langues et des écritures différentes* »[182]. Le projet européen MACS vise à offrir un accès multilingue par sujet aux catalogues bibliographiques des grandes bibliothèques nationales européennes. Dans ce cadre, les utilisateurs pourraient consulter les catalogues des bibliothèques partenaires d'une seule interrogation sujet-matière exprimée dans leur langue maternelle[183]. En effet Jean Delahousse indique que l'alignement entre deux terminologies de langues distinctes, permet de construire un lexique unique décrivant les concepts dans les deux langues, ce qui permettra au « *moteur de recherche [d'effectuer] des recherches sur la base de requêtes dans l'une ou l'autre langue pour arriver aux mêmes ressources documentaires* »[184].

[180]Ibidem.
[181]Voir http://www.rechercheisidore.fr/referentiels
[182]Bourdon Françoise, Boulet Vincent. VIAF : un pivot pour l'accès multilingue à diverses collections.
[183]Voir http://www.*rameau.bnf.fr/informations/pdf/MACS-bnf-2007.pdf*
[184]Delahousse Jean. « Sur l'alignement et la mise en correspondance de terminologies ».

2.3 Porosité des frontières entre production et diffusion : perspectives du web social-sémantique

Il est important de rappeler que les frontières entre la production, la diffusion et la réception s'estompent dans le nouvel environnement du Web qui est également social. Ainsi Gildas Illien indique que « *l'internaute n'est plus en situation de subir ou consommer humblement l'information [mais] il est susceptible d'intervenir dans sa production, son partage et son enrichissement, utilisant les mécanismes du crowdsourcing initiés par le Web 2.0* »[185]. Le rapport du groupe d'incubation du W3C « Bibliothèques et Web de données» exprime ainsi que dans le nouvel environnement du Web social sémantique, les *« ressources peuvent être décrites en collaboration avec d'autres bibliothèques et liées aux données fournies par d'autres communautés, ou même par des individus »*[186].

De nouveaux projets visent justement à travailler sur la réutilisation des tags, des annotations de ressources en langue naturelle ou même sur le texte intégral, afin d'allier les richesses de ce que nous désignons par « web 2.0 » et « web 3.0 ». Par exemple, le projet Faviki[187] est un outil de *bookmarking* social qui permet d'utiliser des concepts Wikipédia, régis par les technologies du web sémantique, comme tags ; le projet OpenCalais[188] vise à extraire des entités nommées de textes en langue naturelle, afin de les aligner ultérieurement avec des référentiels du web de données. Bernard Vatant indique que les applications du web social sont un domaine privilégié et prometteur pour l'application des technologies sémantiques dans les années à venir. Ainsi à propos des technologies du web sémantique, il déclare : « *cette boîte à outils techniques ne présente vraiment d'intérêt que si elle s'accompagne d'une démarche sociale d'ouverture et de partage des données et des savoirs* »[189].

[185]Illien Gildas. « Décrire les objets du savoir, les nouveaux paradigmes du catalogage ».
[186]Groupe d'incubation W3C. Rapport final du groupe d'incubation « Bibliothèques et web de données ».
[187]Voir http://www.faviki.com/pages/welcome/
[188]Voir http://www.opencalais.com/
[189]Vatant Bernard. « Des métadonnées à la description de ressources : les langages du web sémantique »

3. Confiance et gouvernance liées à la création et au maintien des alignements sur le web de données

3.1 Rôle des standards du W3C pour le web sémantique (URI, RDF) et du maintien de l'infrastructure

Nous venons de présenter un écosystème, le web de données, où de nombreux alignements coexistent, qui supposent à la fois une confiance dans le jeu de données auquel on se lie mais aussi une gouvernance pour réaliser et maintenir ces alignements de manière efficiente. La confiance et la gouvernance, comme le rappelle Bernard Vatant, reposent sur l'adoption des standards du web sémantique : d'une part, les URIs pour identifier chaque chose et pour lesquelles la confiance, dans un univers interconnecté et interdépendant, repose sur leur « *stabilité* » et leur « *pérennité* » ; d'autre part, le RDF comme « *langage de description généralisé* »[190]. Une responsabilité accrue incombe alors aux institutions qui publient et sont propriétaires des URIs, car elles doivent intégrer dans leurs missions un *« principe de développement durable de l'économie des connaissances »*[191]. Ainsi Gildas Illien se demande si les agences bibliographiques ne vont pas voir leur rôle évoluer en *« hubs chargés de garantir la maîtrise, la maintenance et la publicité de référentiels de données utiles pour la communauté nationale »*[192] ?

3.2 Rôle des identifiants internationaux, l'exemple d'ISNI

Aussi il est fondamental qu'une gouvernance des identifiants de chaque « chose » sur le web de données culturel soit menée au niveau international. Par exemple, concernant les acteurs « *impliqués dans la création, la production, la gestion et la distribution de contenus intellectuels et artistiques ou faisant l'objet de ces contenus* », ISNI (International Standard Name Identifier) est un « *code international normalisé servant à identifier de manière univoque, sur le long terme et à l'échelle internationale, les personnes et les organismes* »[193]. Ce code est comme l'ISBN[194] pour chaque édition de chaque livre publié, l'ISAN[195] pour les œuvres audiovisuelles ou pour tous les autres identifiants internationaux[196], un moyen efficace de se départir des caractères des différents alphabets, sources de multiples confusions et d'erreurs, pour identifier de manière univoque et pérenne les entités du monde réel à l'échelle mondiale.

[190]Ibidem.
[191]Ibidem.
[192]Illien Gildas. « Décrire les objets du savoir, les nouveaux paradigmes du catalogage ».
[193]Voir http://www.bnf.fr/fr/professionnels/isni_informer/s.isni.html
[194]International Standard Book Number
[195]International Standard Audiovisual Number
[196]e.g. ISRC, ISTC, ISWC, DOI, ISSN, ISMN, EAN, etc

La confiance en cet identifiant repose en grande partie sur la valeur de sa gouvernance, pilotée par l'agence internationale ISNI[197] qui coordonne l'agence d'attribution, l'OCLC[198] (Online Computer Library Center) agréée par l'ISO[199] (Organisation Internationale de Normalisation), une équipe de contrôle qualité constituée d'experts de la British Library et de la Bibliothèque nationale de France, des agences d'enregistrement comme l'agence d'enregistrement ISNI-BnF, ainsi que des membres directs. La base ISNI a été constituée la première fois en 2011, à partir des données de VIAF utilisées comme socle, avant d'être confrontées à des données d'autres contributeurs. Elle a l'objectif d'être une passerelle fiable entre les données de contributeurs de multiples domaines, et favorise ainsi l'interopérabilité dans un écosystème global et mondial des données d'autorité touchant aux identités publiques du domaine culturel.

3.3 Importance de la valeur et de la provenance de l'information et des données

Une autre condition de mise en confiance réside dans la valeur et la provenance de l'information. Ainsi Gildas Illien présente la « *confiance et valeur de l'information* [comme] *deux notions récursives dans le contexte compétitif et chaotique du Web ».* Dans ce contexte les institutions culturelles ont un rôle important à jouer car elles sont à l'origine de « *sources et métadonnées fiables »*[200]. Bernard Vatant nous sensibilise justement sur l'importance de la mémoire et de la citabilité des données sources dans une optique d'évaluation de la valeur de l'information : *« dans le monde ouvert du web, toutes les descriptions ne sont pas équivalentes, et les systèmes d'agrégation de contenu RDF ou les moteurs de recherche devront en tenir compte et, par exemple, mémoriser la provenance des éléments de description qu'ils agrègent »*[201].

3.4 Réseaux et curation de liens, nouveaux enjeux stratégiques et professionnels

De confiance et de gouvernance il est également question avec le choix des jeux de données sur lesquels un organisme va décider de s'aligner. Derrière ces alignements, c'est aussi l'évolution des réseaux qu'il convient d'observer avec vigilance, car « *les nouvelles configurations qui se dessinent au niveau institutionnel et sur le marché mondial sont susceptibles de reconfigurer, voir de déstabiliser, les réseaux existants* »[202]. Nous pouvons étudier ce sujet sous l'angle du concept de la « *curation de données* » qui correspond au fait de « *sélectionner, analyser et réorganiser dans une interface nouvelle des données choisies sur le*

[197]Voir http://www.isni.org/
[198]Voir https://oclc.org/home.en.html
[199]Voir http://www.iso.org/iso/fr/
[200]Groupe d'incubation W3C. Rapport final du groupe d'incubation « Bibliothèques et web de données ».
[201]Vatant Bernard. « Des métadonnées à la description de ressources : les langages du web sémantique »
[202]Illien Gildas. « Décrire les objets du savoir, les nouveaux paradigmes du catalogage ».

web de données »[203]. Ainsi, il s'agit au niveau de la donnée, d'une nouvelle forme de politique documentaire sollicitant les compétences traditionnelles, mais revisitées, des professionnels de l'information. Dans le réseau des bibliothèques, l'expérience OpenCat a mis à jour par exemple, la perspective d'un écosystème où des bibliothèques locales feraient une sélection de données liées d'enrichissement de leurs fonds, depuis une « bibliothèque d'alignements » administrée et maintenue par la Bibliothèque nationale de France[204].

3.5 Enjeux juridiques liés aux données

Le statut juridique des données est également un élément à prendre en compte, car « *l'incitation des pouvoirs publics à adopter des licences ouvertes comme celle recommandée par Etalab*[205] » est un signe fort de la volonté de l'État de mener une politique d'ouverture des données publiques. Ceci peut se résumer par l'engagement de la Ministre de la Culture et de la Communication, Aurélie Filippetti (2012-2014), qui déclare : « *Passer d'une politique de l'accès aux données à une politique de réutilisation des données : c'est le changement de paradigme qui est au cœur de la politique et des usages numériques que je souhaite mener* »[206]. Cependant, selon le principe de l'exception culturelle, il convient de noter que toutes les données culturelles ne disposent pas à ce jour de licences ouvertes.

[203]Bermès Emmanuelle, avec la coll. d'Isaac Antoine et Poupeau Gautier. Le Web sémantique en bibliothèque. p.135
[204]Voir http://www.bnf.fr/fr/professionnels/web_donnees_applications_bnf/a.opencat.html
[205]Voir http://www.etalab.gouv.fr/pages/licence-ouverte-open-licence-5899923.html
[206]Nouvel Observateur, 7 novembre 2013, Entretien d'Aurélie Filippetti, ministre de la Culture et de la Communication, « *Nous devons accompagner les nouveaux usages du numérique* »

C) Bibliographie commentée

Rapports

Groupe d'incubation W3C. *Rapport final du groupe d'incubation « Bibliothèques et web de données »* - Rapport du 25 octobre 2011. trad. (sept. 2012) de « Library Linked Data Incubator Group Final Report ». Consulté le 01/08/2014, en ligne : http://mediatheque.cite-musique.fr/MediaComposite/ARTICLES/W3C/XGR-lld-fr.html

Cette contribution est très importante pour notre rapport de recherche. La plupart de ses thématiques y sont traitées. La description multiple de ressources y est présentée grâce aux URI agissant comme identificateur unique d'une même chose. Est également l'importance de la « *stabilité* » et de la « *fiabilité* » de ces identifiants qui demande un « *cadre de maintenance réglementé en ce qui concerne leurs propriétaires et les responsabilités de ces derniers* ». Les bibliothèques et autres institutions patrimoniales sont présentées comme des acteurs privilégiés pour produire avec pérennité des « *métadonnées fiables* » pour le domaine culturel. Ensuite est évoqué la collaboration distribuée entre les acteurs producteurs de descriptions, à travers la notion de « *déclaration* » d'une ressource, identifiée de manière univoque par une URI. Les déclarations sur une même ressource étant ensuite « *agrégées* » en un « *graphe global* ». Sont présentés également les enjeux professionnels que ce soit au niveau de la production que de la diffusion des données de bibliothèque dans ce nouvel écosystème lié et ouvert. Au niveau de l'exploitation des liens par des artefacts orientés utilisateurs, sont explorés les recherches fédérées plus performantes, les possibilités de navigation enrichies et la facilitation de découvertes heureuses.

Ouvrages

Gandon Fabien, Faron-Zucker Catherine, Corby Olivier. *Le Web sémantique : Comment lier les données et les schémas sur le web ?* Paris : DUNOD, 2012

Cet ouvrage, qui requiert une sensibilité informatique, dresse avec cette approche un portrait introductif et large du web sémantique pour la communauté francophone. Nous nous intéresserons en particulier au sous-chapitre « *Thésaurus : SKOS* » (p. 124-127), qui permet de bien distinguer les vocabulaires de type ontologiques, des vocabulaires utilisés dans les différents Systèmes d'Organisation des Connaissances (thésaurus, taxonomies, systèmes de classification, systèmes d'index) qui sont ceux sur lesquels notre recherche porte. Nous verrons également comment SKOS (Simple Knoweldge Organization System) est un « *vocabulaire RDF qui fournit un modèle commun pour partager et lier sur le Web différents*

Systèmes d'organisation des Connaissances ». Il s'agit ici de la représentation formelle les alignements créés.

Bermès Emmanuelle, avec la coll. d'**Isaac Antoine** et **Poupeau Gautier**. *Le Web sémantique en bibliothèque*. Paris : Editions du Cercle de la Librairie, 2013

Cet ouvrage fait référence au niveau de l'approche du Web sémantique du point de vue des bibliothèques. Il a pour objectif de présenter l'ensemble des enjeux du Web sémantique intéressant les professionnels de cette communauté, afin qu'ils puissent pertinemment appréhender le Web sémantique et mettre en valeurs leurs « *compétences d'analyse de l'information et de traitement des données dans [ce] nouvel environnement* ». Plusieurs chapitres ou sous-chapitres en particuliers intéresserons notre étude, il s'agit de « *Les référentiels clef de voûte du Web de données* » (p. 43-46) et « l'interopérabilité *par les référentiels* » (p. 51-54) qui présentent le rôle de l'alignement de terminologies sur le Web de données, « *La création des liens* » (p. 104-110) qui présente comment les alignements ou mises en correspondance sont réalisés, « *rationaliser la production des données* » (p. 54-57) qui se pose du point de vue de l'exploitation des alignements côté production, et « *Agréger et réutiliser les données* » (p. 133-150) qui se pose du point de vue de l'exploitation des alignements du point de vue de la diffusion, avec notamment l'évocation de la notion de « *curation de données* ».

Blog

Delahousse Jean. « Sur l'alignement et la mise en correspondance de terminologies ». In *Leçons de choses.* Mondeca. Consulté le 01/08/2014, en ligne : http://mondeca.wordpress.com/2009/06/29/sur-l%E2%80%99alignement-et-la-mise-en-correspondance-de-terminologies/

Jean Delahousse est consultant, expert en technologies du Web sémantique. Ce billet est intéressant au niveau de la vision pragmatique et large qu'il offre, car il est le fruit de la mise en place et de la participation à de nombreux projets d'alignements ou de mises en correspondance de terminologies, sur des projets concrets et des travaux de recherche. Après un rappel définitionnel important sur ce que l'on entend par terminologies, alignements et rapprochements de terminologies, est présenté un panachage de nombreux projets ayant eu recours à ce type de stratégie. Ainsi sont exprimés de nombreux besoins conduisant à l'alignement de terminologies du même domaine, la mise en correspondance de terminologies décrivant des domaines différents mais complémentaires ou la mise en relation de

terminologies traitant du même domaine mais sous des angles ou granularité différentes. Ainsi nous pouvons en apprécier la diversité des domaines d'applications : bibliothèques nationales (alignement de RAMEAU avec le thésaurus de la bibliothèque du Congrès), monde de la presse (alignement entre les codes « SLUG » utilisés par les journalistes de l'AFP et la taxonomie internationale IPTC utilisée pour l'annotation des dépêches dans les agences de presse), monde de la santé (mise en relation entre la terminologie médicale Snomed décrivant les maladies avec un référentiel du médicament). Est décrit également dans ce billet « *l'outillage* » nécessaire pour aligner ou mettre en correspondance : Comment on aligne ? Comment on spécifie les relations créées ? Commet on représente et encode l'alignement ou la mise en correspondance ?

Enfin le billet présente une liste intéressante d'exemples d'exploitation des alignements et mises en correspondance, qui peuvent être fait par un système d'information : alimentation des lexiques des moteurs de recherche, enrichissement du lexique d'un moteur de recherche pour traiter des requêtes multilingues, enrichissement de l'annotation des contenus, enrichissement du contenu publié. Nous voyons dans cet article de nombreux cas concrets permettant de mettre en lumière le rôle jouer, ainsi que les conséquences que peuvent apporter les alignements et mises en correspondance de terminologies pour répondre à de nombreux besoins d'interopérabilité entre systèmes d'information.

Articles de périodiques

Isele, Robert et Bizer, Christian. « Active learning of expressive linkage rules using genetic programming ». in *Web Semantics : Science, Services and Agents on the World Wide Web*. Elsevier, 2013

Les auteurs de cette article sont deux chercheurs exerçant dans le groupe de recherche *data and Web science*, de l'université de Mannheim en Allemagne. Ils apportent un éclairage intéressant pour notre recherche sur la partie méthodologique, c'est à dire les règles à suivre, ainsi que les compétences à avoir, pour réaliser des alignements entre entités de bases de données différentes. Ainsi est expliqué qu'il est indispensable d'avoir une bonne connaissance de la structure des données que l'on veut aligner. Ensuite, il est nécessaire d'effectuer les bons choix, que ce soit par rapport aux propriétés discriminantes des entités que l'on va sélectionner et comparer, tout comme par rapport à de la distance de mesure que l'on va mobiliser ainsi que le seuil de distance qui va être spécifier pour aligner. L'article présente également l'étape nécessaire de « *normalisation* » des chaînes de caractères avant de procéder à la comparaison. L'étape de comparaison consiste elle en une évaluation de la « *similarité* »

de valeurs, suivant des mesures de distance (« *specific distance mesure* ») tels que *Levenshtein*, *Jaccard* ou la distance géographique. Il est également présenté l'intérêt et la nécessité de devoir combiner dans certains cas plusieurs comparaisons de propriétés pour pouvoir identifier et rapprocher des entités (ex : pour différencier Berlin en Allemagne de Berlin aux États-Unis). L'article met en lumière l'intérêt un outil d'alignement de type semi-automatique, « *active learning* », qui a pour objectif de proposer des listes réduites de candidats à l'alignement qui seront soumis à la validation d'un usager.

Briatte Katell. « Hadoc, un programme pour harmoniser les données culturelles ». In Documentaliste - Sciences de l'information. Vol. 51, n2, 2014

Katell Briatte chef de projet en maîtrise d'ouvrage du programme Hadoc, au Ministère de la Culture et de la Communication. Cet article nous présente les problématiques de description des ressources culturelles aussi bien du point de vue de la production que de celui aujourd'hui très fortement lié de la diffusion des données. En effet, si sur le Web, les usagers souhaitent accéder aux contenus de manière unifiée et cohérente, il n'en reste que la diffusion large des données, est tributaire des pratiques de production des descriptions de ressources des divers objets culturels, pratiques résultant de « *processus métiers répondant à des objectifs particuliers suivant des méthodologies propres encadrées par des normes spécifiques* ».

Pour répondre à cette problématique, il est intéressant de constater que le programme Hadoc (Harmonisation de la production des données culturelles) a pour ambition d'accroître la qualité des données le plus en amont du cycle de vie des données, c'est à dire dès leur production. C'est ainsi que le programme vise à harmoniser les processus de productions des données de l'ensemble des acteurs métier de la description de ressources culturelles, que ce soit au niveau du modèle de données, « *l'ossature* », que de référentiels terminologiques, « *la chair* ». Alors si les silos techniques sont rompus, ce sont également les métiers et organisations qui se trouvent décloisonnés.

La réflexion autour de l'harmonisation des données d'autorités, au sein d'un écosystème Web contributif et sémantique, ouvre des pistes intéressantes sur la mutualisation des charges de travail, ainsi que sur la complémentarité et la complétude des données entre institutions.

L'ambition ultime du programme Hadoc, est de fournir un référentiel des biens culturels, où une « *carte d'identité* » unique de chacun de ces biens culturels, partagée entre tous les processus métiers, permettra de garantir l'authenticité, l'unicité, la qualité, l'exhaustivité et la pérennité de données de référence, auprès des producteurs comme des usagers.

Illien Gildas. « Décrire les objets du savoir, les nouveaux paradigmes du catalogage ». In *Les métiers de l'information et la « donnée » : analyse d'un monde en mutation*. Documentaliste – Sciences de l'information. N3 sept. 2013

Cet article nous apporte des informations sur les notions de « *confiance* » ainsi que de « *valeur* » de l'information dans le contexte de l'abondance aussi « *compétitive* » que « *chaotique* » de l'information numérique sur le Web, c'est deux notions liées réciproquement. Également est évoqué la conduite du changement dans ce nouvel écosystème, avec pour corollaire la dimension économique qui dans un contexte où les bibliothèques sont soumises à d'importantes contraintes budgétaires, conduira à « *réaliser des économies d'échelle en recherchant des solutions de mutualisation* ». Ainsi les institutions ont une nécessité économique à ne plus produire seules leurs métadonnées et doivent s'inscrire dans des réseaux à des fins de « *partage* », d' « *échange* » ou d' « *enrichissement* ». L'auteur indique également que les agences bibliographiques voir leurs rôles évoluer, en prenant la direction de « *hubs chargés de garantir la maîtrise la maintenance et la publicité de référentiels de données utiles pour la communauté nationale* ». Est mentionné également les bienfaits de la mutualisation des descriptions de ressources, dans la perspective offerte à chaque institution, de se concentrer et de porter son expertise là où elle a une « *valeur ajoutée propre* ». Enfin, les questions soulevées par l'inscription d'un établissement au sein de réseaux nous montre de manière significative comment les problématiques techniques autour de la création de liens, débouchent sur des problématique très stratégiques comme le choix des bons réseaux répondant à ses attentes, que ce soit au niveau institutionnel, national, européen ou international.

Bermès Emmanuelle. « Bibliothèques, archives et musées : l'enjeu de la convergence des données du patrimoine culturel. »

Emmanuelle Bermès est conservateur des bibliothèques. Elle a notamment travaillé sur l'évolution des catalogues de la BnF vers le Web sémantique (projet data.bnf.fr) ainsi que sur le projet *linked data* du Centre Pompidou Virtuel. L'article présente les enjeux autour de la convergence des donnés du patrimoine culturel. Un premier constat important côté usager est que du point de vue de l'internaute, sa recherche d'information n'est « *pas centrée sur les institutions mais sur les contenus* ». Ainsi les données de bibliothèques, d'archives et de musées sont au cœur de cette volonté de convergence. Mais l'interopérabilité et l'ouverture concerne également « *d'autres domaines métiers différents, et notamment d'acteurs*

commerciaux ». L'auteur pointe les limites des approches traditionnelles d'interopérabilité des données patrimoniales, du point de vue de l'usager (pauvreté des résultats de recherche et non présence sur le parcours naturel de l'internaute, des portails d'accès fédéré à plusieurs bases de données) comme pour l'utilisation de protocoles pas en adéquation avec les autres acteurs du Web (Z39.50, OAI-PMH).

L'auteur présente ensuite les avantages de l'interopérabilité sur le Web de données grâce aux standards du Web. Est mis en avant les notions d'accès et de navigation sur le Web de documents, notions qui sont déclinées au niveau du Web de données. Ainsi l'enjeu est alors de lier et de donner accès aux données de bibliothèques, de musées et d'archives, sans que l'usager est connaissance des problématiques liées aux bases de données et aux formats hétérogènes entre ces données. C'est un élément important abordé, sur « l'opacité » pour un usager de l'architecture technique d'un site construit avec les technologies du Web sémantique. Dans cet article, les différences de modèles conceptuels entre les bibliothèques (concepts de notices bibliographiques et notices d'autorité, migration vers les FRBR), les musées (Cidoc CRM : le concept d'événement est central) et les archives (notion de contexte et de hiérarchie) nous sont bien présentés, tout comme la possibilité sur le Web de données de « *créer des liens entre des ressources décrites suivant divers modèles* » (grâce à la grammaire commune RDF et le recours aux *mappings* avec les vocabulaires ou ontologies du Web sémantique) sans essayer de les réduire à un modèle commun appauvrissant les particularités de chacun. Il est enfin exposé le rôle prépondérant des référentiels sur le Web de données, « *en particulier lorsqu'il s'agit de construire l'interopérabilité entre des données de domaines différents* ».

Zacklad Manuel. « Évaluation des systèmes d'organisation des connaissances » in *Les cahiers du numérique,* 2010/3 Vol. 6, p. 133-166

Cet article nous permet de bien identifier les différents systèmes d'organisations des connaissances (SOC), en les replaçant dans une perspective historique, et en en comprenant les propriétés et la complémentarité. Cette analyse se positionne du point de vue à la fois des usages comme de la conception de ces SOC, en prenant en compte les évolutions liées à l'émergence du web sémantique et socio sémantique.

Compte rendus de Séminaires, conférences

Vatant Bernard. « Des métadonnées à la description de ressources : les langages du web sémantique ». In *Métadonnées : mutations et perspectives. Séminaire INRIA 29 septembre-3 octobre 2008 – Dijon*. ADBS, 2008

Ce compte rendu de séminaire nous livre des éléments intéressants concernant le sujet de la gouvernance sur le Web de données. En effet, il nous est expliqué l'importance et la responsabilité que représentent la publication et le maintien d'URI « *stables* » et « *pérennes* ». Il est également présenté la question de l'enjeu de la « *provenance* » des données. En effet, dans un environnement de description relié et distribué, ces éléments sont indispensables pour garantir un niveau de « *confiance* » nécessaire au bon fonctionnement de ce système.

Bourdon Françoise, Boulet Vincent. *VIAF : un pivot pour l'accès multilingue à diverses collections*. IFLA 2011. Consulté le 26/08/2014 [en ligne] : http://conference.ifla.org/past-wlic/2011/79-bourdon-fr.pdf

Cet article permet de mesurer les caractéristiques de valeur, de fiabilité et de confiance que présentent les données d'autorité de bibliothèques sur le web de données. Les auteurs démontrent l'intérêt de l'alignement dans VIAF, qui permet de créer de l'interopérabilité entre ressources tout en conservant les particularités locales de catalogage et d'écriture.

Isaac Antoine. « Les référentiels : typologie et interopérabilité ». *In Le document numérique à l'heure du web de données. Séminaire IST INRIA 2012.* Consulté le 27/08/2014 [en ligne] : http://hal.inria.fr/docs/00/84/38/41/PDF/isaac-v2.pdf

Cet article nous permet de bien analyser la notion de référentiel, notamment dans le cadre d'une approche *linked data*. L'auteur traite des questions d'alignement entre vocabulaires de valeur et systèmes d'organisation des connaissances. Il pointe alors les perspectives vertueuses de « *liens entre les référentiels des systèmes documentaires traditionnels* [avec d'] *autres systèmes à base de connaissances* » dont les données sont publiées sur le web de données (e.g. DBpedia, GeoNames). Est introduit aussi la notion de « jeux de données de référence » qui renvoie au rôle de confiance et de fiabilité que peuvent jouer les grandes institutions, pour servir de point d'ancrage à d'autres jeux de données sur le web de données.

Dalbin Sylvie. « Représentation et accès à l'information : transformation à l'œuvre ». in *Métadonnées : mutations et perspectives. Séminaire INRIA 29 septembre-3 octobre 2008 – Dijon*. ADBS éditions, 2008.

Cette contribution permet de bien cerner l'ensemble du processus de création et d'administration des métadonnées, de la conception métier (modélisation de la réalité et formalisation du modèle conceptuel, développement de référentiels métiers) à la conception informatique (modèle de données informatique, schéma de données et encodage). L'auteur évoque également la notion de « *contrainte de l'existant* » dans les nombreux projets liés au développement de systèmes d'information. Le contexte de création des listes d'autorité est également bien présenté.

D) Conclusion

Ce travail de recherche a permis de faire un point terminologique sur ce que l'on entend par alignements de vocabulaires d'autorités sur le web de données culturel. Ensuite il s'est intéressé d'une part aux méthodes et outils utilisés pour implémenter ces alignements, connaissance indispensable pour en comprendre les implications technologiques, professionnels et stratégiques. Les algorithmes de rapprochement de jeux de données redessinent en effet les rapports entre l'humain et la machine dans la gestion des métadonnées culturelles. L'infrastructure du web de données culturel demande une gouvernance forte afin de recréer de la confiance dans un environnement pris entre continuité et changement. Cette gouvernance repose aussi bien sur le recours commun aux standards du W3C pour le web sémantique, notamment les URIs et le RDF, que sur la maintenance de ces URIs et de leur infrastructure sur le web. Elle nécessite également le recours à des outils normalisés pour identifier les instances des entités culturelles à l'échelle internationale (e.g. ISNI) et redessine par ailleurs les réseaux aux échelles nationale et internationale, où les frontières traditionnelles s'estompent pour faire place à de nouveaux partenariats entre les différents acteurs de la culture (bibliothèques, archives, musées) et même, au-delà de ce cercle, entre des acteurs du monde culturel et des acteurs venus d'autres horizons (e.g. Wikipédia).

D'autre part, le travail de recherche s'est tourné sur l'intérêt profond de ces alignements entre vocabulaires d'autorité. S'intéresser à l'exploitation de ces alignements aussi bien du point de vue de la production que de celui de la diffusion des données a permis d'éclairer les enjeux couvrant l'ensemble du cycle d'administration des données. Cela nous a permis de comprendre l'intérêt qu'ont les institutions à mutualiser leur effort de production de métadonnées : elles peuvent en tirer des bénéfices en termes aussi bien économiques que qualitatifs (dans un scénario où chacun se spécialiserait sur sa valeur ajoutée propre). Du point de vue des utilisateurs, et en particulier des internautes, nous avons pu explorer comment ces alignements pouvaient rentrer en adéquation avec leurs pratiques, que ce soit par la présence sur le

parcours (« *flow* ») de l'utilisateur (e.g. alignement sur Wikipédia[207]), la navigation inter-institutionnelle centrée sur le contenu et ainsi la sérendipité favorisée, ou encore les possibilités de recherche augmentée pour interroger des moteurs de recherche avec des index enrichis et multilingues.

Ce travail nous a permis d'appréhender l'ampleur de l'écosystème qui se met en place autour de l'alignement de vocabulaires d'autorité sur le web de données culturel. Nous comprenons qu'il se situe entre la continuité et des changements parfois profonds, un bon exemple étant les possibilités entrevues par le web social-sémantique où l'utilisateur devient également producteur dans un cadre beaucoup plus contrôlé que ne l'était la folksonomie[208] (recours à des autorités sur le web de données alignées sur les tags pour indexer).

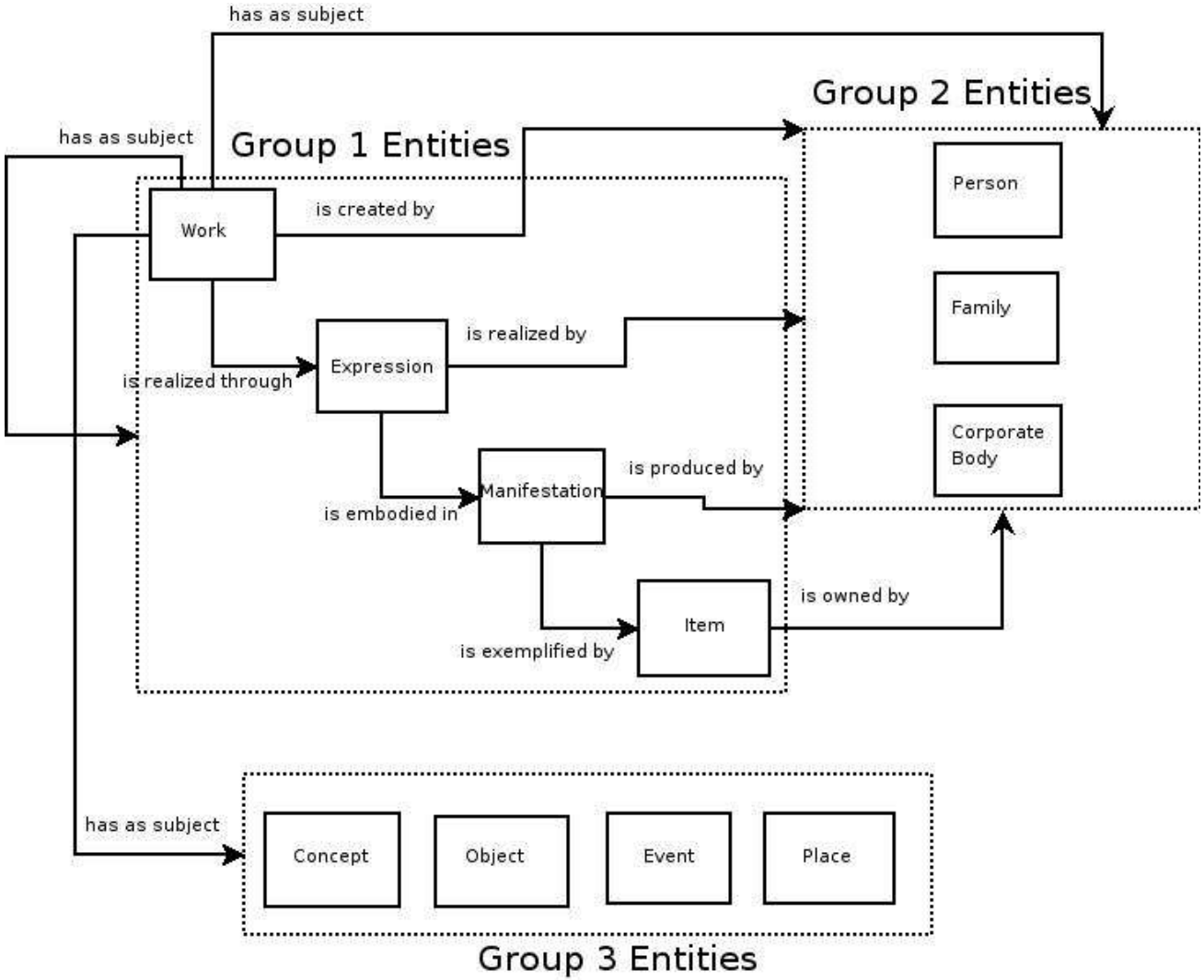
Cependant, de nombreux terrains non explorés dans ce travail restent à étudier, comme par exemple la problématique des partenariats entre les secteurs publics et privés du domaine culturel. Nous pensons par exemple à la collaboration entre les éditeurs, les distributeurs et les bibliothèques dans l'interprofession du livre. D'autres alignements sont encore en gestation ou à inventer, afin de fournir davantage de richesse informationnelle et de services innovants aux utilisateurs. Une veille approfondie sur les projets de Recherche et Développement actuels serait très utile. Enfin, cet écosystème ouvert et lié, où chaque créateur peut être identifié et rattaché à toutes ses œuvres et autres « *records* » en lien avec ses activités, suscite des interrogations de nature sociétale comme par exemple la protection des individus et des données personnelles dans ce nouvel espace d'information. Ces aspects pourraient faire l'objet de travaux complémentaires pour approfondir un sujet aux multiples facettes.

[207] Voir liens data.bnf.fr dans la page Auguste Bailly de Wikipédia : http://fr.wikipedia.org/wiki/Auguste_Bailly
[208] Voir http://bbf.enssib.fr/consulter/bbf-2006-04-0066-002

ANNEXES

Annexe I : présentation des entités et leurs relations du modèle FRBR

(*source* : Groupe de travail IFLA. Fonctionnalités requises des notices bibliographiques : rapport final : http://www.bnf.fr/documents/frbr_rapport_final.pdf)



Annexe II : mapping formats originaux / vocabulaires RDF

- Aurorités titres (MARC > RDF)

Libellé catalogue	Zone intermarc	Unimarc	Correspondance RDF
Lieu de création	008 position 12-13	102 $a	rdagroup1elements:placeOfOriginOfTheWork
Langue	008 position 14-16	101 $a	dcterms:language
Date de création	008 position 27-36		dcterms:date
Auteur principal	100, 110	240 $a,	dcterms:creator
Forme retenue (TUT)	141	230, 730	skos:prefLabel@in_lang (type= skos:concept)
Forme retenue (TUM)	144	230, 240, 730, 740	skos:prefLabel@in_lang (type= skos:concept)
Forme retenue (TIC)	145	230, 240, 730, 740	skos:prefLabel@in_lang (type= skos:concept)
Formes rejetées	441, 444, 445	430, 440	skos:altLabel @in_lang
Lien vers une notice spécifique d'œuvre	302 $3	530 $3 $5h, 540	ore:aggregates
entre une oeuvre (anonyme ou non) et un auteur personne physique ou	321 $3, 322 $3	500, 510, 520	dcterms:contributor
Lien générique vers une autre œuvre	502 $3	530 $3 $5g, 540	ore:isAggregatedBy
Note d'information publique	600 $a	305ab, 310ab,	dcterms:description
Note publique sur les sources consultées avec	610 $a	810 $a	skos:editorialNote

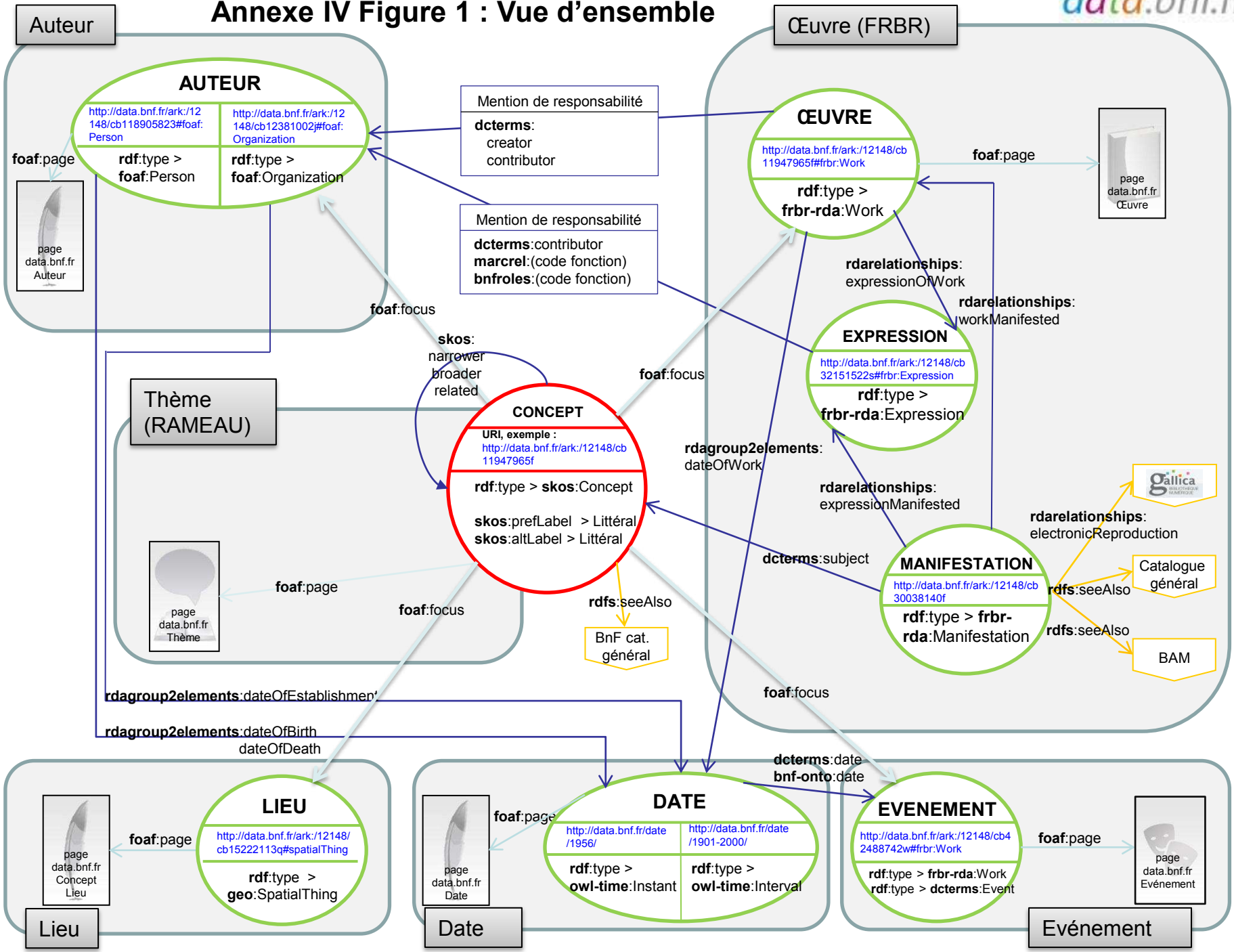
- Titre archives et manuscrits (EAD > RDF)

Libéllé catalogue	Élément EAD dans sa hiérarchie	Correspondance RDF
Titre d'œuvre	<ead><archdesc><did><unittitle>	dc:title
Intitulé archivistique	<ead><archdesc><did><unittitle>	dc:title
Cote	<ead><archdesc><did><unitid	bnf-onto:cote
Date	<ead><archdesc><did><unitdate	dc:date
Importance matérielle	<ead><archdesc><did><physdesc	dc:description
Document numérisé	<ead><archdesc><dao	rdarelationships:electronicRepro
Role	<persname role=""> ou <famname role=""> ou <corpname role=""> ou	bnfrole

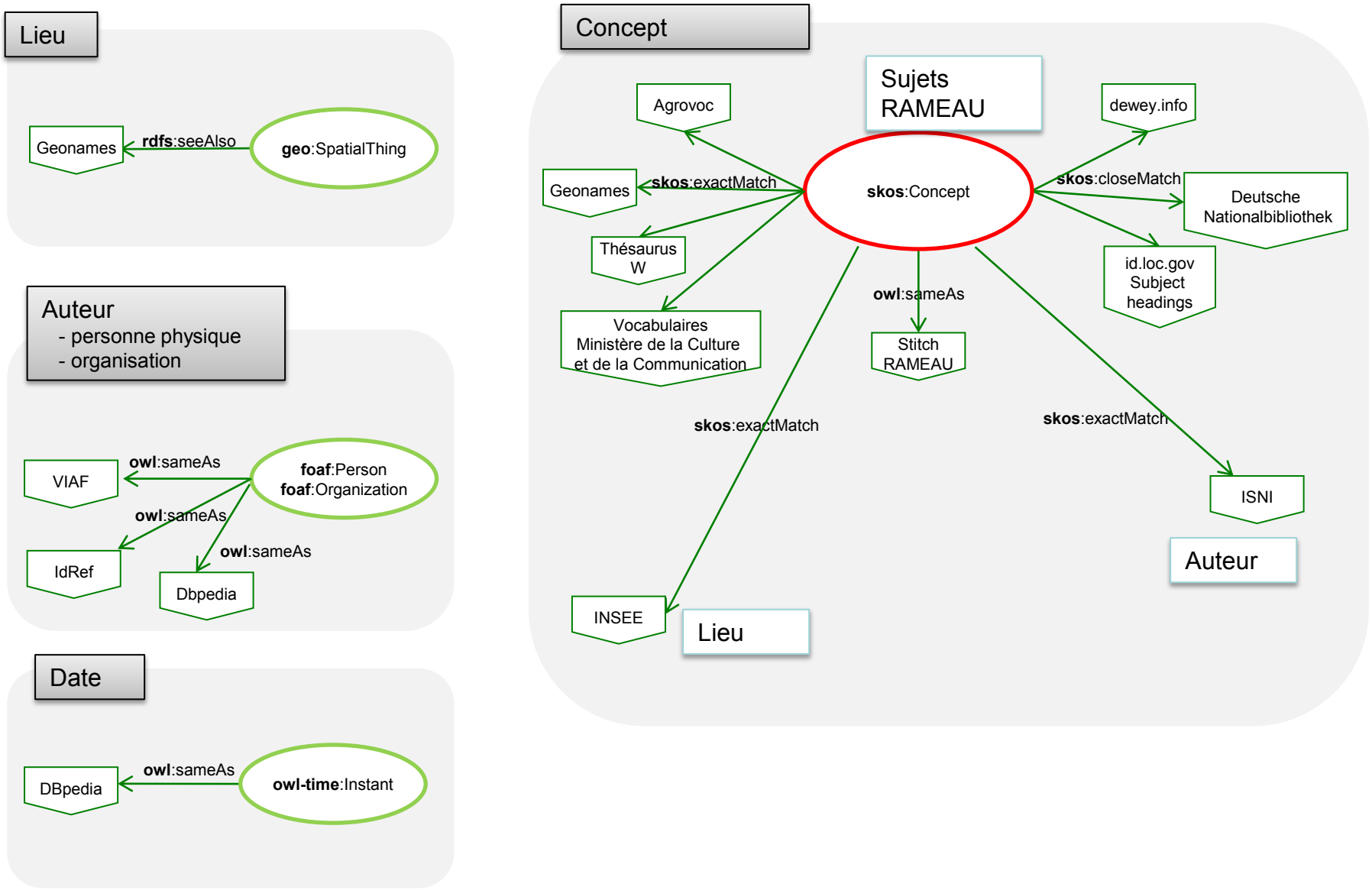
Annexe III : de la notice à la donnée (exemple de l'autorité Victor Hugo)

Annexe IV Figure 1 : Vue d'ensemble

data.bnf.fr



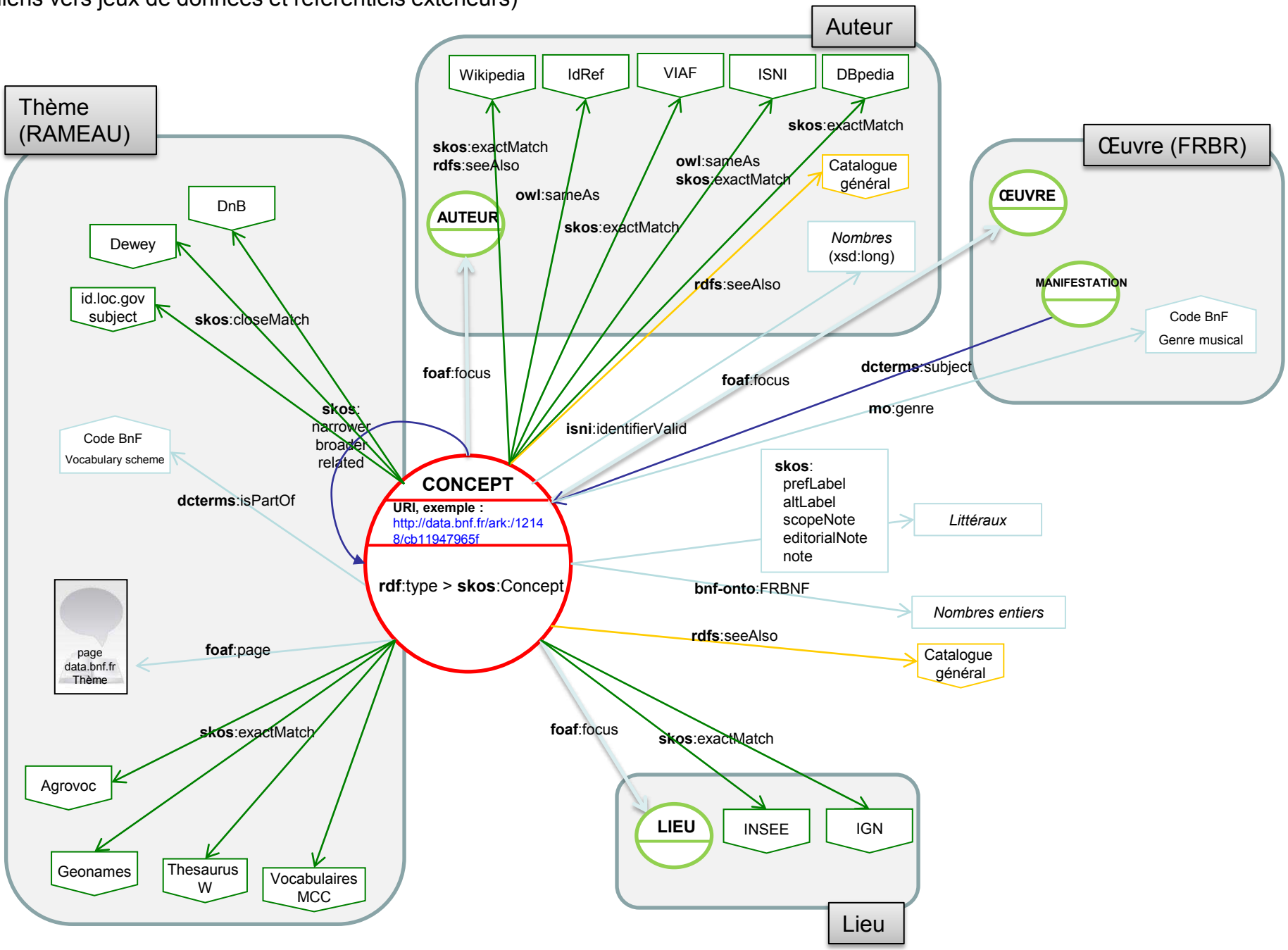
Annexe IV Figure 2 : Alignements vers des jeux de données externes



Annexe IV figure 3 : Focus sur le Concept

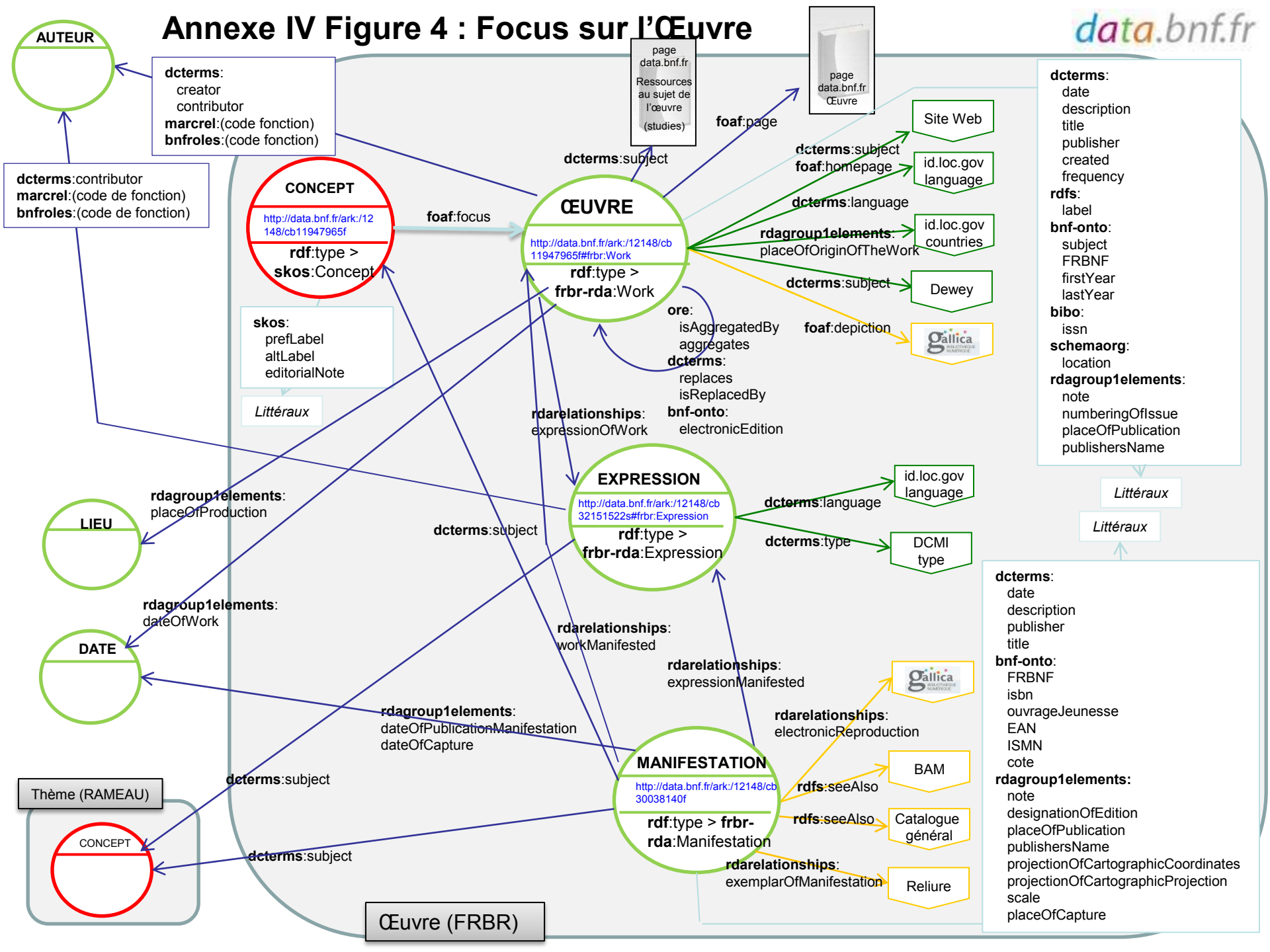
(liens vers jeux de données et référentiels extérieurs)

data.bnf.fr

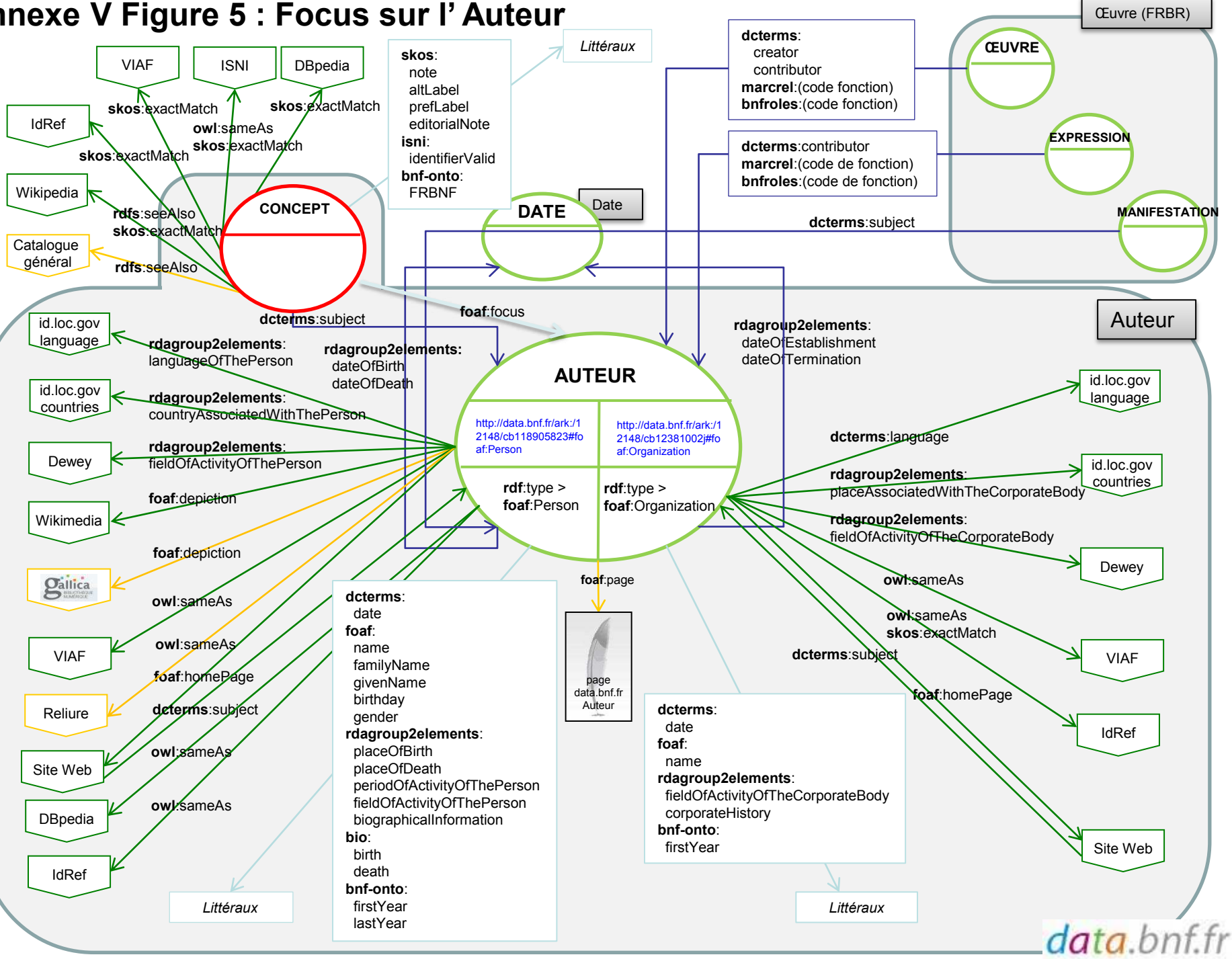


Annexe IV Figure 4 : Focus sur l'Œuvre

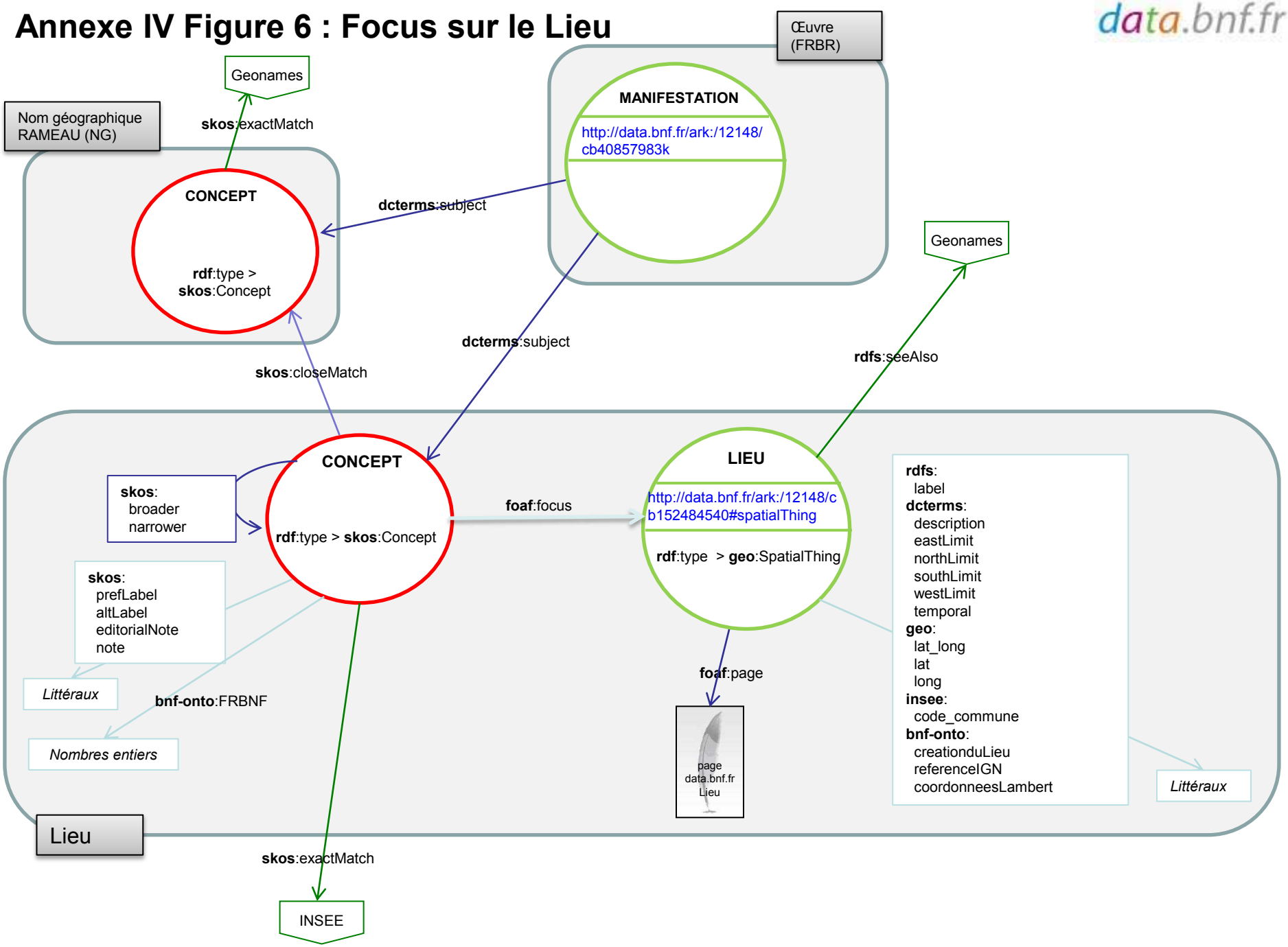
data.bnf.fr



Annexe V Figure 5 : Focus sur l' Auteur



Annexe IV Figure 6 : Focus sur le Lieu



Annexe IV Figure 7 : Focus sur la Date

data.bnf.fr

Œuvre (FRBR)

ŒUVRE

MANIFESTATION

Auteur

AUTEUR

rdf:type > foaf:Person

rdf:type > foaf:Organization

Littéraux

rdfs:label

rdagroup2elements: dateOfBirth dateOfDeath

rdagroup2elements: dateOfEstablishment

rdagroup1elements: dateOfWork

page data.bnf.fr Date

foaf:page

rdagroup2elements: dateOfTermination

DATE

http://data.bnf.fr/date/1956/

http://data.bnf.fr/date/1901-2000/

rdf:type > owl-time:Instant

rdf:type > owl-time:Interval

rdagroup1elements: dateOfCapture

rdagroup2elements: dateOfPublicationManifestation

owl:sameAs

DBpedia

dcterms:hasPart

dcterms:date

bnf-onto:date

Evénement

EVENEMENT

Date

Annexe V : requêtes types SPARQL pour interroger le *triplestore* de data.bnf.fr

1.1 Découvrir le modèle de données

Quelles sont les propriétés pour décrire la ressource http://data.bnf.fr/ark:/12148/cb11933798p

```
SELECT DISTINCT ?p ?o WHERE {
<http://data.bnf.fr/ark:/12148/cb11933798p> ?p ?o.
}
```

1.2 Informations sur une personne ou une organisation

1.2.1 Informations sur l'auteur

Exemple : Toutes les dates biographiques des auteurs (anniversaire et année de naissance et de mort) ainsi que leur nom,

```
SELECT ?auteur ?jour ?date1 ?date2  ?nom where {
?auteur foaf:birthday ?jour.
?auteur bio:birth ?date1.
?auteur bio:death ?date2.
OPTIONAL {?auteur foaf:name ?nom}
}
ORDER BY (?jour)
```

Exemple 2 : toutes les informations sur Victor Hugo et toutes les formes du nom de cette personne

```
SELECT DISTINCT ?nom_complet ?nom ?prenom ?forme_retenue
?formes_rejetees ?pays ?langue ?sexe ?anniversaire ?date_naissance
?lieu_naissance ?date_mort ?lieu_mort ?periode_activite ?domaine_activite
?bio WHERE {
<http://data.bnf.fr/ark:/12148/cb11907966z#foaf:Person> foaf:gender ?sexe;
<http://rdvocab.info/ElementsGr2/countryAssociatedWithThePerson>
?pays;
<http://rdvocab.info/ElementsGr2/languageOfThePerson> ?langue;
<http://rdvocab.info/ElementsGr2/dateOfBirth> ?date_naissance;
<http://rdvocab.info/ElementsGr2/placeOfBirth> ?lieu_naissance;
foaf:birthday ?anniversaire;
<http://rdvocab.info/ElementsGr2/dateOfDeath> ?date_mort;
<http://rdvocab.info/ElementsGr2/placeOfDeath> ?lieu_mort;
<http://rdvocab.info/ElementsGr2/fieldOfActivityOfThePerson>
?domaine_activite;
<http://rdvocab.info/ElementsGr2/biographicalInformation> ?bio;
foaf:name ?nom_complet;
foaf:familyName ?nom;
foaf:givenName ?prenom.
<http://data.bnf.fr/ark:/12148/cb11907966z> skos:altLabel ?formes_rejetees;
skos:prefLabel ?forme_retenue
OPTIONAL {<http://data.bnf.fr/ark:/12148/cb11907966z#foaf:Person>
<http://rdvocab.info/ElementsGr2/periodOfActivityOfThePerson>
?periode_activite}}
```

Exemple 3 : L'anniversaire des compositeurs, avec leurs noms et leurs années de naissance et de mort

```
SELECT ?auteur ?jour ?date1 ?date2  ?nom where {
?auteur foaf:birthday ?jour.
?doc bnfroles:r220 ?auteur.
?auteur <http://rdvocab.info/ElementsGr2/dateOfBirth> ?date1.
?auteur <http://rdvocab.info/ElementsGr2/dateOfDeath> ?date2.
OPTIONAL {?auteur foaf:name ?nom}
}
ORDER BY (?jour)
```

Exemple 4 : retrouver, pour chaque auteur, les pages comprenant les documents au sujet de cet auteur

```
SELECT DISTINCT ?auteur ?documents_a_propos_de WHERE {
?doc dcterms:creator ?uri_auteur .
?documents_a_propos_de dcterms:subject ?uri_auteur .
?uri_auteur foaf:name ?auteur
FILTER REGEX (?documents_a_propos_de, ".*studies.*")
}
```

**Cette requête renvoie uniquement des pages de data.bnf.fr (qui comprennent 'studies' dans leur URI)*

```
SELECT DISTINCT ?auteur ?documents_a_propos_de WHERE {
?doc dcterms:creator ?uri_auteur .
?documents_a_propos_de dcterms:subject ?uri_auteur .
?uri_auteur foaf:name ?auteur
FILTER (!REGEX (?documents_a_propos_de, "http://data.bnf.fr/.*"))
}
```

**À l'inverse, cette requête renvoie uniquement les pages extérieures à data.bnf*

Exemple 5 : retrouver, pour un auteur (ici Charles Baudelaire), les documents au sujet de cet auteur

```
SELECT DISTINCT ?document WHERE {
?document dcterms:subject <http://data.bnf.fr/ark:/12148/cb118905823> .
}
```

1.2.2 Liens entre auteurs et œuvres

Exemple : retrouver toutes les œuvres de Victor Hugo avec les titres

```
SELECT DISTINCT ?work ?title WHERE
{
<http://data.bnf.fr/ark:/12148/cb11907966z> foaf:focus ?person.
?work dcterms:creator ?person ;
      rdfs:label ?title .
}
```

1.3 Information sur une œuvre

Exemple 1 : Retrouver tous les titres du Roman de la Rose et la note associée

```
SELECT DISTINCT ?titre_forme_internationale_francais ?formes_rejetees ?note_associee
WHERE {
<http://data.bnf.fr/ark:/12148/cb166125510> skos:altLabel ?formes_rejetees;
skos:prefLabel ?titre_forme_internationale_francais;
skos:editorialNote ?note_associee
}
```

Exemple 2 : retrouver la liste des éditions d'une même œuvre (ici, les Fleurs du mal)

```
SELECT DISTINCT ?title ?date ?editeur WHERE {
<http://data.bnf.fr/ark:/12148/cb11947965f> foaf:focus ?Work .
?Work rdarelationships:expressionOfWork ?Expression .
?Manif     rdarelationships:expressionManifested ?Expression.
OPTIONAL{?Manif dcterms:date ?date}
OPTIONAL{?Manif dcterms:title ?title}
OPTIONAL{?Manif dcterms:publisher ?editeur}
}
```

1.4 Liens entre auteurs et expressions : fonctions d'un auteur sur des documents

Exemple 1 : Trouver les traducteurs d'ouvrages italiens en français

```
SELECT DISTINCT ?notice ?name WHERE {
?Work dcterms:language <http://id.loc.gov/vocabulary/iso639-2/ita> .
?Exp bnfroles:r680 ?traducteur ;
    dcterms:language <http://id.loc.gov/vocabulary/iso639-2/fre>.
?Work rdarelationships:expressionOfWork ?Exp .
?traducteur foaf:name ?name.
?notice foaf:focus ?traducteur.
}
```

Exemple 2 : Retrouver le nom et prénom de tous les photographes

```
SELECT DISTINCT ?Photographe ?Prenom ?Nom WHERE
{
?expression bnfroles:r530 ?Photographe .
?Photographe a foaf:Person.
OPTIONAL {?Photographe foaf:givenName ?Prenom.}
OPTIONAL {?Photographe foaf:familyName ?Nom.}
}
```

1.5 Sujets (RAMEAU)

Exemple 1 : Pour un sujet Rameau, récupérer tous les termes spécifiques de niveau 1 et, le cas échéant, de niveau 2

```
SELECT DISTINCT ?original_rameau ?prefLabel ?uri_a ?label_a ?uri_b ?label_b
WHERE {
?original_rameau skos:prefLabel ?prefLabel ;
        skos:narrower ?uri_a .
MINUS {?original_rameau foaf:focus ?focus .}
?uri_a skos:prefLabel ?label_a .
OPTIONAL {
?uri_a skos:narrower ?uri_b .
?uri_b skos:prefLabel ?label_b .
}
}
```

Exemple 2 : Documents au sujet du thème « Escrime » (URI *http://data.bnf.fr/ark:/12148/cb11931273g*)

```
SELECT ?doc {
?doc dcterms:subject <http://data.bnf.fr/ark:/12148/cb11931273g>.
}
```

Exemple : Les éditions qui ont pour sujet la photographie

```
SELECT DISTINCT ?Edition ?title WHERE {
?Edition a frbr-rda:Manifestation;
            dcterms:subject <http://data.bnf.fr/ark:/12148/cb11933113t>;
            dcterms:title ?title.
}
```

1.6 Identifiants

Exemple 1 : Pour un ISNI donné, récupérer le nom et prénom de la personne concernée (ici Alexandre Dumas)

```
SELECT DISTINCT ?nom ?prenom  WHERE
{
?person isni:identifierValid "0000000121012885" ;
          foaf:focus ?identity.
          ?identity foaf:familyName ?nom;
          foaf:givenName ?prenom.
}
```

Exemple 2 : Trouver l'identifiant ARK d'une notice à partir de son numéro FRBNF (ici, œuvre « les travailleurs de la mer »

```
SELECT DISTINCT ?idArk WHERE
{
?idArk bnf-onto:FRBNF "11992081"^^xsd:integer.
}
```

Exemple 3 : Retrouver la ou les œuvres correspondant à un ISBN avec le nom de l'auteur

Propriété utilisée : bnf-onto:isbn
Exemple : Trouver « Extension du domaine de la lutte » à partir de l'ISBN 2-7028-4777-3.

```
SELECT DISTINCT ?work ?title ?name WHERE
{
?work rdfs:label ?title;
      dcterms:creator ?creator.
?work rdarelationships:expressionOfWork ?expression.
?manifestation rdarelationships:expressionManifested ?expression.
?manifestation bnf-onto:isbn "2-7028-4777-3".
?creator foaf:name ?name.}
```

1.7 Vocabulaires et référentiels : genres musicaux, langue, pays, type de document, ouvrages jeunesse

Exemple 1 : sélectionner tous les documents de type image

```
SELECT ?image where {
?image dc:type <http://purl.org/dc/dcmitype/StillImage>.
} Limit 100
```

Exemple 2 : Lister les œuvres musicales par genre, en commençant par les genres les plus représentés

```
SELECT DISTINCT ?genre ?label COUNT (?work) AS ?nbWork
WHERE {
?work mo:genre ?genre.
?genre skos:prefLabel ?label }
ORDER BY DESC (?nbWork)
```

Exemple 3 : Sélectionner tous les ouvrages adaptés pour la jeunesse et l'œuvre adaptée correspondante

```
SELECT ?uri ?oeuvre WHERE {
?manifestation bnf-onto:ouvrageJeunesse "true"^^xsd:boolean ;
        rdarelationships:expressionManifested ?expression ;
        rdfs:seeAlso ?uri.
?oeuvre rdarelationships:expressionOfWork ?expression.
}
```

1.8 Requête sur des dates

Jour de naissance (anniversaire) sous la forme mm-jj : foaf:birthday
La date exacte de naissance ou de mort sous la forme aaaa-mm-jj : bio:Birth et bio:Death
Les années de naissance et de mort uniquement : bnf-onto:firstYear et bnf-onto:last year

Exemple : liste des auteurs morts avant 1924

```
select distinct ?nom ?auteur ?mort where {
?oeuvre dcterms:creator ?auteur .
?auteur rdf:type foaf:Person ;
bio:Death ?mort ;
foaf:name ?nom
FILTER (xsd:integer (?mort) < "1924"^^xsd:integer )
}
ORDER BY DESC (?mort)
```

*ne ramène pas les dates qui ne correspondent pas à un nombre entier (e.g. les dates incertaines n'apparaissent pas ici

Liste des auteurs nés avant 1500, triés par date de naissance

```
SELECT ?auteur ?naissance where {?auteur bnf-onto:firstYear ?naissance.
?auteur a foaf:Person.
Filter (?naissance < "1500"^^xsd:integer)
}
ORDER BY ASC (?naissance)
```

1.9 Lieux, notices géographiques

Retrouver tous les documents numérisés au sujet des notices géographiques

```
SELECT DISTINCT ?lieu ?doc ?docnum where {
?lieu rdf:type geo:SpatialThing.
?doc dcterms:subject ?concept.
?concept foaf:focus ?lieu.
?doc rdarelationships:electronicReproduction ?docnum.
}
```

Retrouver tous les documents numérisés au sujet des notices géographiques et des thèmes associés

```
SELECT DISTINCT ?lieu ?doc ?docnum where {
{
```

```
?lieu rdf:type <http://www.w3.org/2003/01/geo/wgs84_pos#SpatialThing>.
?doc dcterms:subject ?concept.
?concept foaf:focus ?lieu.
?doc rdarelationships:electronicReproduction ?docnum.}
UNION {
?doc dcterms:subject ?rameau.
?lieu skos:exactMatch ?rameau.}
}
```

Retrouver les lieux alignés avec leurs thèmes correspondants :

```
SELECT ?lieu ?concept where {
?c a <http://www.w3.org/2003/01/geo/wgs84_pos#SpatialThing>. ?lieu foaf:focus ?c. ?lieu
skos:closeMatch ?concept.}
```

1.10 Spectacles

1.11 Liens vers des documents numérisés dans Gallica

Exemple 1 : retrouver tous les documents numérisés sur un thème

Exemple : Les photographies (rôle photographe) numérisées dans Gallica, avec leur URL et leur titre (1000 premiers résultats)

```
SELECT DISTINCT ?Doc ?title ?URLGallica WHERE {
?Expression bnfroles:r530 ?Creator.
?Doc rdarelationships:expressionManifested ?Expression.
?Photo rdarelationships:electronicReproduction ?URLGallica; dcterms:title
?title.
}
```

1.12 Illustration d'une page

Exemple 1 : Retrouver les illustrations proposées pour la page Charles Baudelaire

```
SELECT DISTINCT ?image WHERE
{
<http://data.bnf.fr/ark:/12148/cb118905823#foaf:Person>      foaf:depiction
?image.
}
```

Exemple 2 : Images de Gallica de type « portraits » et les identifiants ISNI correspondants, pour les auteurs français

```
select ?auteur ?url  where {
?auteur rdf:type skos:Concept ;
foaf:focus ?person.
?doc rdarelationships:electronicReproduction ?url.
?doc dcterms:subject ?auteur ;
dcterms:subject <http://data.bnf.fr/ark:/12148/cb11932843f>.
?person                <http://rdvocab.info/ElementsGr2/countryAssociatedWithThePerson>
<http://id.loc.gov/vocabulary/countries/fr#>.}
```

Annexe VI : tableau de comparaison des schémas de métadonnées employées pour les référentiels Rameau Noms Géographiques et Géo

	Géo			Rameau		
Forme retenue	170	**$a**	**Elément d'entrée**	167	**$a**	**Elément d'entrée**
		$o	**Inversion**		**$o**	**Inversion**
		$c	Localisation			
		$g	Désignation		$g	Qualificatif
					$x	Subdivision de sujet ou de forme
					$y	Subdivision géographique
					$z	Subdivision chronologique
Formes rejetées	470	$a	Elément d'entrée	467	$a	Elément d'entrée
		$o	Inversion		$o	Inversion
		$c	Localisation			
		$g	Désignation		$g	Précision

Annexe VII : spécifications pour l'amélioration de l'algorithme d'alignement Géo / Rameau NG

Principes : aller chercher des alignements qui ne sont pas faits grâce à des règles métiers. Garder en vue que les algorithmes doivent traiter de gros volumes de données et que les cas particuliers pouvant être traités à la main ne rentrent pas dans son périmètre d'action. Egalement faire attention à ne pas produire de règles moins conservatrices et riches en alignement, mais produisant de nombreux faux positifs.

Sommaire

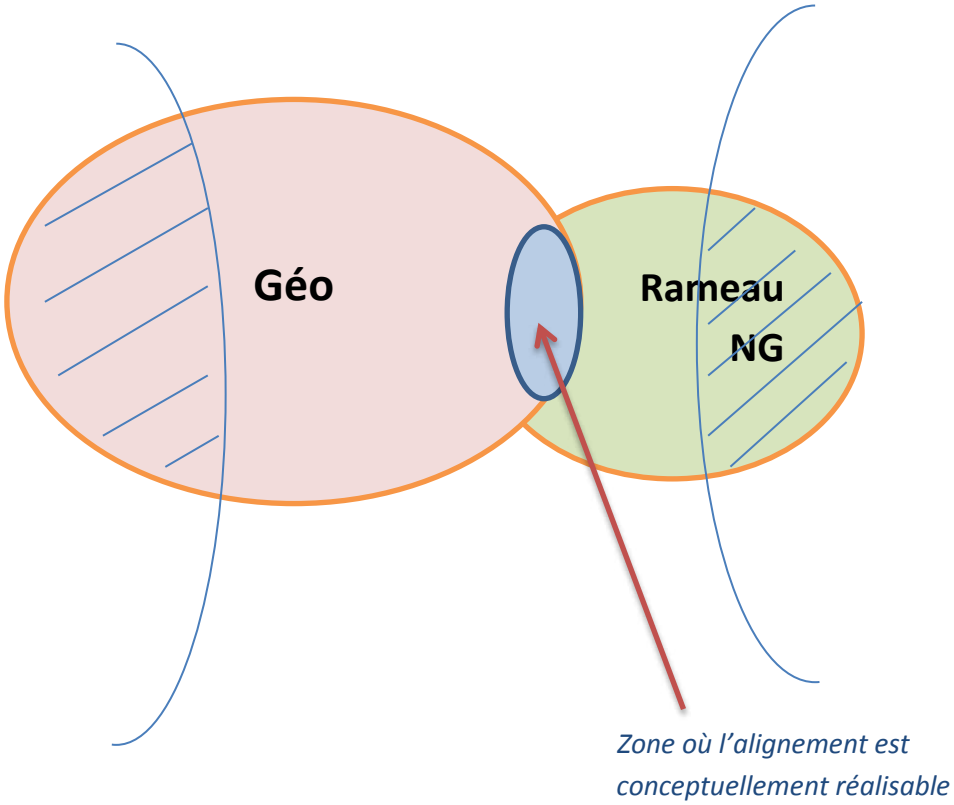
1) Ne garder que les candidats

Objectif : alléger les calculs en vue des algorithmes, exclure ce qui n'a aucune chance de s'aligner et réduire le risque de « faux positifs » sur des algorithmes d'alignements moins conservatifs.

Constat : toutes les Rameau NG et les Géo n'ont pas nécessairement leurs équivalents. Il est possible par une règle conservatrice de ne garder que les candidates et d'éliminer ceux qui n'ont aucune chance de s'aligner.

2 règles identiques :

- Une Géo qui ne partage dans son élément d'entrée [$a] de ses formes retenues et rejetées (zones 170 et 470) aucun mot commun avec l'élément d'entrée [$a] d'une Rameau NG des formes retenues et rejetées (zones 167 et 467) ne peut avoir d'alignement Rameau correspondant car celui-ci n'existe pas alors conceptuellement,
- Une Rameau NG qui ne partage dans son élément d'entrée [$a] de ses formes retenues et rejetées (zones 167 et 467) aucun mot commun avec l'élément d'entrée [$a] des formes retenues et rejetées (zones 170 et 470) d'une Géo ne peut avoir d'alignement Géo correspondant car celui-ci n'existe pas alors conceptuellement. La Rameau NG en question ne peut alors être candidate à l'alignement.



2) Deux grands cas métiers identifiés

Cas n°1 : à l'intérieur des villes

Exemple : rue, quartier, grotte, place, château, temple, chapelle…

A : îlotisation (blocage) de localisation géographique sur la ville

Eléments à rapprocher :

Zone(s) Géo	170 $a
Zone(s) Rameau NG	167 $a

Après normalisation (accent, casse), algorithme de rapprochement préconisé : exactMatch

+ Désambiguïsation par la région puis le pays quand cela est possible

- Coté Géo zone 170 : $c (région), $c (pays)
- Coté Rameau zone 167 : $g (pays)

B : Alignement sur le spécifique de la ville

Eléments à rapprocher (il s'agit des qualificatifs utilisés dans les formes retenues et rejetées de chaque référentiel pour décrire ce type d'information) :

Zone(s) Géo	170 $b* 470 $b*
Zone(s) Rameau NG	167 $x 467 $x

*(*le $b peut se répéter)*

Après normalisation (accent, casse), algorithme de rapprochement préconisé : exactMatch.

Alignement de deux entrées si un match et un seul se produit ; sinon listing pour réalisation de l'alignement par un opérateur

Cas n°2 : les « territoires » et les « accidents géographiques »

Exemples de territoires : département, région, canton, empire, principauté, district…

Exemples d'accidents géographiques : cours d'eau, massif, estuaire, lac, courant, péninsule…

Contraintes :

- l'indexation Rameau utilise le $g (en 170 / 470) à des fins de localisation (ex : une région ou un pays) aussi bien qu'à des fins de qualification typologique sur les territoires (ex : « département ») ou les accidents géographiques (ex : « estuaire »).
- les pratiques d'indexation sont assez variables (du côté Rameau comme du côté Géo), c.a.d parfois il y a utilisation du $g (désignation / qualificatif) et parfois utilisation du $o (élément du nom inversé) pour marquer ce type d'information.

Exemples Rameau :

167 $w....b.....$aSeine$oEstuaire de la$gFrance

167 $w....b.....$aGironde$gFrance$gestuaire

Exemples Géo :

170 $w.1..b.fre.$aCarlingford Lough$cLouth$cIrlande$gestuaire

170 $w 1 b fre $aSeine$oEstuaire de la$cFrance

Constats :

- Il apparait nécessaire de distinguer dans le $g des Rameau les qualificatifs typologiques (territoires, accidents géographiques) de ceux de localisation (région / pays). Ceci dans le but de faire travailler exclusivement un algorithme de rapprochement sur ces données isolées.
- Il apparait nécessaire de croiser les $g et $o pour trouver des alignements

A : Préparation :

Ne conserver dans le $g des Rameau NG que les qualificatifs de type « territoires » ou « accidents géographiques ».

Une solution proposée : s'appuyer sur les termes d'une liste stable et finie, pour « nourrir » l'algorithme (voir *liste-termes-blocage-territoires-accidentsgeo.txt*[1]), et exclure du corpus tous les autres mots se trouvant dans le $g de Rameau ne figurant pas dans cette liste. A noter que cette liste est très représentative mais pas exhaustive (pour cela il faudrait faire une extraction de la zone concernée en DPI).

[1] La liste comprend les mots suivants : Cours eau;Royaume;mont;Massif;Château;Région;Département;Canton;République;détroit;Province;Etat;Comté;Duché;
Site archéologique;volcan;Ville ancienne;mer;col;empire;principauté;district;estuaire;courant;péninsule;océan;monts;plateau;mont; île; vallée; îles; lac; bassin; forêt; plaine;

B : îlotisation (blocage) contextuel : « territoires / accidents géo »

Eléments à prendre en compte : (il s'agit des zones de marquage des territoires et accidents géographiques dans les zones de qualificatifs ou d'inversion, des formes préférées et rejetées)

Zone(s) Géo	170 $g 170 $o* 470 $g 470 $o*
Zone(s) Rameau NG	167 $g 167 $o* 467 $g 467 $o*

Après simplification (accent, casse et stopwords), algorithme de rapprochement préconisé : exactMatch.

**attention, y a-t-il un risque avec le $o qu'il matche des choses d'un autre ordre que les territoires et accidents géographiques ?*

Si tel était le cas, nous pourrions envisager par exemple d'exclure ces mots sources d'erreurs, ou encore se baser sur une liste destinée à nourrir l'algorithme (après extraction des occurrences du $o côté géo et rameau).

C : Alignement sur le libellé

Eléments à rapprocher : (il s'agit des libellés des formes préférées et rejetées)

Zone(s) Géo	170 $a 470 $a*
Zone(s) Rameau NG	167 $a 467 $a*

*(*à tester si l'on veut être moins conservatif et réaliser plus d'alignements)*

Après normalisation (accent, casse), algorithme de rapprochement préconisé : exactMatch.

+ Désambiguïsation par la région ou le pays quand cela est possible ($c côté Géo et $g côté Rameau)

Alignement de deux entrées si un match et un seul se produit ; sinon listing pour réalisation de l'alignement par un opérateur

Annexe VIII : alignement de « FRBRisation » du catalogue