Université de Montréal

Le Web de données et le Web sémantique à Bibliothèque et Archives nationales du Québec : constats et recommandations fondés sur l'initiative de la Bibliothèque nationale de France

par Marielle St-Germain

École de bibliothéconomie et des sciences de l'information

Faculté des arts et sciences

Mémoire présenté à la Faculté des études supérieures et postdoctorales

en vue de l'obtention du grade de maître

en Sciences de l'information

Mai 2016

Résumé

Ce mémoire traite des concepts et de l'implantation du Web sémantique et du Web de données au sein d'institutions documentaires. Une analyse et une définition des technologies caractérisant ces concepts sont d'abord présentées dans l'objectif de les clarifier et d'assurer une bonne compréhension des différents enjeux qui en découlent pour les acteurs du domaine. Ensuite, les éléments démontrant la pertinence et les défis pour les professionnels de l'information sont décrits. Puis, l'objectif est d'analyser le processus de mise sur pied d'un projet de Web de données au sein de la Bibliothèque nationale de France pour proposer une transposition possible au contexte de Bibliothèque et Archives nationales du Québec, en vue d'une application. La liste des treize étapes pour l'implantation d'un projet de Web de données en bibliothèque ainsi que la proposition de l'application d'une méthodologie de développement de logiciel à ces pratiques sont ensuite présentées. Suite à cette analyse, des recommandations quant aux différentes étapes d'implantation sont proposées.

Mots-clés : Web sémantique, Web de données, Web 3.0, Bibliothèques, BAnQ, BnF, Recommandations

Abstract

This dissertation discusses the concepts and implementation of Semantic Web and Linked Data within libraries. Analysis and definition of technologies characterizing these concepts are first presented with the objective to clarify and ensure a good understanding of the various issues arising for actors in the field. Then, the elements demonstrating the relevance and challenges for information professionals are described. The objective is to analyze the implementation process of a Linked Data project within the Bibliothèque nationale de France to propose a possible transposition to the context of Bibliothèque et Archives nationales du Québec, for an application within the latter. A list of thirteen steps for the implementation of a library Linked Data project and the proposal for applying a software development process on these practices are presented. Following this analysis, recommendations regarding these various stages of implementation are proposed.

Keywords : Semantic Web, Linked Data, Libraries, BAnQ, BnF, Recommendations

Table des matières

Liste des figures

Liste des tableaux

Liste des sigles et acronymes

AACR – *Anglo-American Cataloguing Rules*

ACFAS – Association francophone pour le savoir

AIP – *Archival Information Package*

ANQ – Archives nationales du Québec

API – Interface de programmation applicative

ARK – *Archival Resource Key*

ASCII – *American Standard Code for Information Interchange*

ABAnQ – Les Amis de Bibliothèque et Archives nationales du Québec

BAC – Bibliothèque et Archives Canada

BAnQ – Bibliothèque et Archives nationales du Québec

BIBFRAME – *Bibliographic Framework*

BIBO – *Bibliographic Ontology*

BnF – Bibliothèque nationale de France

BNQ – Bibliothèque nationale du Québec

BQ – Bibliographie du Québec

CBQ – Catalogue des bibliothèques du Québec

CDWA – *Categories for the description of works of art*

CIDOC – Comité international pour la documentation

CIDOC CRM – *CIDOC Conceptual Reference Model*

DCMI – *Dublin Core Metadata Initiative*

DIP – *Dissemination Information Package*

DNF – Dépôt numérique fiable

DOI – *Digital Object Identifier*

DOL – Données ouvertes et liées

EAN – *European Article Numbering*

EDM – *Europeana Data Model*

FOAF – *Friend of a friend*

FRAD – *Functional Requirements for Authority Data*

FRBR – *Functional Requirements for Bibliographic Records*

FRSAD – *Functional Requirements for Subject Authority Data*

HTML – *Hypertext Markup Language*

HTTP – *HyperText Transfer Protocol*

IdO – Internet des objets

IFLA – Fédération internationale des associations et institutions de bibliothèques

Inria – Institut national de recherche en informatique et en automatique

Insee – Institut national de la statistique et des études économiques

IoT – *Internet of Things*

IRI – *Internationalized Resource Identifier*

ISBD – *International Standard Bibliographic Description*

ISBN – *International Standard Book Number*

ISMN – *International Standard Music Number*

ISSN – *International Standard Serial Number*

JSON – *JavaScript Object Notation*

JSON-LD – *JavaScript Object Notation for Linked Data*

LC – *Library of Congress*

LCSH – *Library of Congress Subject Headings*

MARC – *Machine-Readable Cataloging*

MCC – Ministère de la Culture et des Communications

METS – *Metadata Encoding and Transmission Standard*

OAI-ORE – *Open Archives Initiative Object Reuse and Exchange*

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

OAIS – *Open Archival Information System*

OCLC – *Online Computer Library Center*

ODbL – *Open Database License*

OWL – *Web Ontology Language*

PEB – Prêt entre bibliothèques

Pistard – Programme informatisé servant au traitement des archives et à la recherche documentaire

PREMIS – *PREservation Metadata : Implementation Strategies*

RAMEAU – Répertoire d'autorité-matière encyclopédique et alphabétique unifié

RDA – *Resource Description and Access*

RDF – *Resource Description Framework*

RDFa – *Resource Description Framework in Attributes*

RDFS – *Resource Description Framework Schema*

ReLIRE – Registre des Livres Indisponibles en Réédition Électronique

REST – *Representational State Transfer*

RFN – Réseau francophone numérique

RIF – *Rule Interchange Format*

RUP – *Rational Unified Process*

RVM – Répertoire de vedettes-matières

SIGB – Système intégré de gestion de bibliothèque

SIP – *Submission Information Package*

SKOS – *Simple Knowledge Organization System*

SPAR – Système de préservation et d'archivage réparti

SPARQL – *SPARQL Protocol and RDF Query Language*

SPIRL – *Stanford Prize for Innovation in Research Libraries*

SQL – *Structured Query Language*

SQTD – Service québécois de traitement documentaire

SWEO – *Semantic Web Education and Outreach*

Turtle – *Terse RDF Triple Language*

UCS – *Universal Coded Character Set*

UML – Langage de modélisation unifié

UNESCO – Organisation des Nations unies pour l'éducation, la science et la culture

URI – *Uniform Resource Identifier*

URL – *Uniform Resource Locator*

URN – *Uniform Resource Name*

VIAF – *Virtual International Authority File*

VRA – *Visual Resource Association*

W3C – *World Wide Web Consortium*

WoT – *Web of Things*

XML – *Extensible Markup Language*

À mes parents.

Remerciements

Ce mémoire a été réalisé grâce à l'aide et au soutien de plusieurs à qui j'aimerais adresser mes sincères remerciements. D'abord, merci à ma directrice de recherche, Madame Lyne Da Sylva, pour ses conseils, sa patience et le temps qu'elle m'a consacrés durant tout ce processus d'apprentissage. Elle a su me guider et m'appuyer dans ma démarche tout en me transmettant sa passion pour la recherche. Ce fut un honneur et un plaisir d'être sous sa direction. J'aimerais ensuite remercier Monsieur Jean-François Gauvin pour sa confiance, son appui et d'avoir partagé avec moi son expertise. Je remercie aussi Mesdames Raphaëlle Lapôtre et Agnès Simon d'avoir répondu à mes questions avec honnêteté et précision.

J'aimerais aussi saluer ma famille et mes amis, pour leur compréhension et leurs encouragements durant ce cheminement. Merci de croire en moi et de votre soutien constant.

1. Introduction

Le domaine des sciences de l'information a la caractéristique de regrouper en son sein plusieurs champs de connaissance particuliers qui ont chacun leurs propres spécificités. La bibliothéconomie et l'archivistique ont donné naissance à des techniques fiables et reconnues visant l'organisation, l'évaluation, la collecte, la préservation ainsi que la diffusion d'informations de toute nature, qu'il s'agisse de documents physiques ou numériques, mais aussi des données qui s'y rattachent. Les changements qui gravitent autour des formats de documents et des moyens de les partager et les diffuser créent une nécessité de repenser les pratiques et modèles.

L'objectif général de ce mémoire est d'explorer un ensemble de technologies spécifiques au Web sémantique et au Web de données, puis de proposer une liste de treize étapes pour la mise sur pied de projets mettant de l'avant ces technologies au sein d'institutions documentaires. Cette liste vise à faciliter l'appropriation par les professionnels de l'information.

1.1 Contexte

Pour les professionnels de l'information, la nécessité de s'adapter aux changements technologiques et aux nouvelles pratiques des utilisateurs en matière de recherche d'information n'est pas inconnue. De nouveaux formats pour véhiculer et stocker les connaissances ont vu et continuent à voir le jour, ce qui complique, dans une certaine mesure, le travail de gestion et de diffusion des savoirs, mais aussi crée un besoin de repenser les pratiques et standards. L'évolution des métiers en bibliothèques fut inévitable suite à la montée grandissante des médias de masse qui ont changé les habitudes de lecture et de consommation de la population (Turner, Dufour, Laplante, Leroux et Salaün, 2009). Cependant, l'arrivée du Web, et du numérique par le fait même, a non seulement modifié les comportements informationnels, mais aussi notre façon de consommer et de traiter les connaissances ainsi que nos interactions. De plus, ces changements sont survenus brusquement, en un peu moins d'une vingtaine d'années. La

multiplication des formats de documents a eu pour conséquence directe la prolifération de données les décrivant, nécessitant de nouvelles compétences et connaissances de la part des professionnels de l'information (Stuart, 2011). Malgré cette réalité, la raison d'être et les missions des bibliothèques sont restées les mêmes : offrir l'accès à une collection de documents pour sa communauté, acquérir des ressources ainsi que produire des services et, finalement, agir comme un intermédiaire entre l'usager et les ressources (Leroux et al., 2009). On comprend aujourd'hui l'importance que peuvent prendre les données relatives aux ressources numériques et, peu à peu, on évalue la possibilité de les utiliser afin de répondre à ces missions en saisissant le potentiel d'utilisation qu'elles présentent.

1.1.1 Impacts du World Wide Web

Traditionnellement, le rôle du bibliothécaire était celui d'acquérir, de traiter, d'organiser et de préserver des documents imprimés et d'aider les usagers à localiser l'information recherchée (Rao et Babu, 2001; Stuart, 2011). Avec l'arrivée du Web, la numérisation des documents, l'apparition des catalogues de bibliothèques informatisés et des moteurs de recherche, on a vu la perception du rôle de bibliothécaire changer. En effet, les tâches ont été redéfinies au fil des ans pour laisser plus de place aux conseils et à l'aide à la recherche ainsi qu'à une collaboration entre professionnels de l'information et techniciens informatiques. Les qualités reconnues aux professionnels d'information telles que l'importance accordée aux besoins des usagers et au partage de connaissances, la capacité d'identifier et d'organiser des documents imprimés et numérisés ainsi que la connaissance des concepts relatifs à la gestion leur ont permis de comprendre les avantages qui découlaient de l'arrivée du Web (Rao et Babu, 2001; Stuart, 2011). Cependant, on reconnait une certaine menace en ce qui a trait à la survie de la profession (Hicks, 2013) : vivrons-nous éventuellement dans un monde complètement numérique ? La profession de bibliothécaire survivra-t-elle à ces nouvelles technologies ? De plus, les bibliothécaires manquent de connaissances informatiques et doivent faire face à la croyance selon laquelle il est possible de tout trouver sur Internet par soi-même.

1.1.2 Impacts du Web 2.0

L'arrivée du Web 2.0, vers la fin des années 2000, s'est traduite elle aussi par un besoin d'adaptation de la part des bibliothèques. Il est primordial de noter que le Web n'est pas un phénomène statique, mais toujours en évolution, ce qui fait en sorte que tous les domaines se doivent de s'adapter et de prévoir que le rôle joué aujourd'hui ou hier ne sera pas nécessairement celui de demain (Stuart, 2011). Plusieurs définitions du Web 2.0 ont été proposées, la première tentative provenant Tim O'Reilly (2005). Celui-ci propose sept principes permettant de caractériser le concept :

- Le Web en tant que plateforme ;
    - Le Web permet la création d'applications et ne se limite plus uniquement à la propagation d'information. On constate qu'il s'agit d'une plate-forme misant sur le partage.
- Exploiter l'intelligence collective ;
    - La participation des utilisateurs permet une démocratisation du Web. Ceux-ci peuvent contribuer au contenu de différents sites Web, que ce soit de par ses connaissances (p. ex. Wikipédia), des évaluations (p. ex. Amazon), des commentaires ou des recommandations. Plus les utilisateurs contribuent, plus il y a de valeur à un site Web.
- La puissance se trouve dans les données ;
    - La valeur d'une application dans le Web 2.0 est influencée par les données qu'elle rend accessible et qu'elle entretient. Par exemple, plusieurs grandes compagnies ouvrent leurs données, permettant aux développeurs de créer des applications tierces et Google récupère les données recueillies lors des recherches effectuées par les utilisateurs.
- La fin des cycles de *release* (changements de versions) ;
    - Contrairement aux logiciels traditionnels, les utilisateurs n'ont pas besoin d'attendre les différents changements de versions. Les applications évoluent constamment et les changements aux versions sont disponibles en ligne, s'opérant souvent de manière automatique.

- Les modèles de programmation légers ;
    - Les outils et les données utilisés pour programmer les applications sont faciles d'utilisation.
- Les logiciels ne dépendent plus d'un seul appareil (l'ordinateur personnel) ;
    - Les applications sont maintenant accessibles sur un grand nombre de plateformes différentes telles que les consoles de jeux vidéo, les téléphones intelligents, les lecteurs MP3, etc.
- L'enrichissement de l'expérience utilisateur.
    - L'utilisation de technologies telles que AJAX ou JavaScript permettent la création d'interfaces utilisateur plus riches et facilitant la navigation.

Stephens et Collins (2007), pour leur part, indiquent que le Web 2.0 est caractérisé par l'utilisation d'outils numériques qui permettent aux usagers de créer, modifier et publier du contenu de toute sorte. L'accent est alors mis sur les médias sociaux. Ils présentent les principes qui définissent cette nouvelle utilisation du Web :

- Conversations ;
    - La participation des usagers, la discussion et la rétroaction sont encouragées.
- Communauté ;
    - Les conversations peuvent mener à un sentiment d'appartenance à une communauté sur un réseau social.
- Participation ;
    - De l'information nouvelle est créée grâce à la collaboration entre les usagers et tout le monde peut créer du contenu qui peut être réutilisé et modifié gratuitement.
- Expérience ;
    - L'engagement au sein de cette communauté est enrichissant et peut mener à un sentiment de réalisation de soi.
- Partage.
    - Les usagers peuvent ou non discuter de leur vie personnelle.

Les bibliothèques ont su rapidement s'approprier les outils de cette nouvelle plateforme, tels que Facebook, Twitter, YouTube, Flickr, en y assurant une présence continue (Farkas, 2007). Ces tribunes gratuites donnent la possibilité d'obtenir une meilleure visibilité des différents services offerts, d'éduquer les usagers aux techniques utilisées (p. ex., la numérisation d'un ouvrage) et d'interagir avec eux. La question de l'importance et de l'efficacité de la présence des bibliothèques sur les médias sociaux a été largement étudiée (Aharony, 2009; Bradley, 2015; Koontz et Mon, 2014; Liew, Wellington, Oliver et Perkins, 2015; Stuart, 2011; Swanson, 2012; Thomsett-Scott, 2014). Grâce à cet outil, une information partagée par un usager peut rejoindre tout son réseau de contacts en peu de temps, ce qui permet une plus grande diffusion de l'information.

1.1.3 Potentiel du Web sémantique

Le Web sémantique, pour sa part, propose une nouvelle manière de traiter de façon automatique l'information et les données qui la constituent sur le Web. Présentement, bien qu'elles puissent être comprises par les humains, les informations que l'on retrouve sur les pages Web ne sont pas structurées de façon à ce qu'elles puissent être traitées et comprises par les logiciels, dans la mesure où elles n'ont pas de valeur sémantique. L'objectif du Web sémantique est donc de permettre, grâce à des technologies spécifiques menant à cette « compréhension » des données par les machines, de trouver, partager, réutiliser ou modifier de l'information de manière plus rapide et efficace. Pour ce faire, le *World Wide Web Consortium* (W3C) travaille à la mise sur pied de standards qui permettent l'identification univoque et pérenne des entités (noms de personnes, concepts, lieux, événements, sujets, etc.), l'expression des relations sémantiques entre ces entités et des raisonnements complexes sur les données. Le Web sémantique est particulièrement intéressant pour le domaine des sciences de l'information, car il permet de faciliter la recherche d'information, d'améliorer l'accessibilité aux données structurées et d'augmenter la visibilité des institutions et de leurs collections sur le Web.

1.2 Problématique

En 2005, l'Organisation des Nations unies pour l'éducation, la science et la culture (UNESCO) publie un rapport intitulé « Vers les sociétés du savoir » qui propose une définition de la « société de la connaissance » ainsi qu'une vision du futur et des défis qui s'y rattachent. L'UNESCO met l'accent sur les changements qu'a apportés la troisième révolution industrielle, c'est-à-dire celle des nouvelles technologies, et sur les différences entre la société de l'information et la société de la connaissance qui en découlent ainsi que l'importance de ne pas les confondre. Bien que les deux concepts soient directement liés, la société de l'information est caractérisée par les progrès technologiques constatés alors que la société de la connaissance fait référence aux aspects sociaux, éthiques et politiques qui sont associés à ces mêmes progrès. Pour l'UNESCO, la société de la connaissance repose sur l'accès universel aux savoirs, la liberté d'expression ainsi que l'innovation et la création de nouveaux modèles de développement. Son objectif est de favoriser l'autonomisation et d'encourager le développement humain en se basant sur les droits de l'homme, une importante diffusion des informations et une accumulation des connaissances. Il est intéressant de souligner à quel point ces thèmes se rapprochent des missions fondamentales et des objectifs des bibliothèques publiques présentées dans le manifeste de la Fédération internationale des associations et institutions de bibliothèques (IFLA) et de l'UNESCO sur la bibliothèque publique (1994). En effet, le manifeste souligne l'importance que les services et collections soient accessibles à tous et qu'ils ne soient pas l'objet de censure idéologique, politique ou religieuse ou de pressions commerciales. De plus, on y indique la nécessité qu'ont les bibliothèques publiques d'être « (…) à la fois reflet des tendances du moment et de l'évolution de la société, et mémoire de l'entreprise et de l'imagination humaines. » La démocratisation de la connaissance, vue comme un bien public, est au cœur de ces missions, aux côtés de l'importance accordée à sa diffusion et son partage. On constate donc la possibilité pour les professionnels de l'information de devenir des acteurs majeurs au sein de cette société de la connaissance.

Plusieurs enjeux relatifs à la société de la connaissance sont cependant soulevés par l'UNESCO. D'abord, la fracture numérique et la fracture cognitive qui l'accompagne présentent

un obstacle à l'accessibilité aux connaissances. Selon le *World Development Report 2016 : Digital Dividends,* publié par la Banque mondiale en 2016, 60 % de la population mondiale n'a pas accès au Web et ne peut donc pas participer aux activités économiques qui l'entourent. Toujours selon ce rapport, l'Internet favorise les opérations économiques, mais ne s'y limite pas. En effet, le Web permettrait une meilleure participation des femmes à la vie active, une plus grande facilité à communiquer pour les personnes handicapées et la création d'emplois. Ensuite, un second enjeu mis de l'avant par l'UNESCO est celui de la nécessité de faire des mises à jour périodiques dues à la vitesse de l'innovation technologique. Cet aspect peut être une source de difficultés dans certains milieux dans la mesure où l'arrivée de nouveaux modèles d'apprentissage et la formation continue nécessitent de nombreuses ressources et peuvent parfois faire l'objet d'une certaine réticence face au changement. Troisièmement, le rapport souligne le fait que les nouvelles technologies et le numérique ont eu comme conséquence une prolifération de l'information. Cependant, une importante quantité, voire un excès, d'information ne signifie pas pour autant une abondance de savoirs.

En effet, il est nécessaire de développer nos capacités cognitives et notre esprit critique pour bien évaluer la qualité de l'information à laquelle nous sommes exposés. On se doit donc, comme communauté mondiale, d'assurer qu'un nombre suffisant d'individus soient aptes à analyser, trier et considérer l'information pour, par la suite, l'insérer dans des bases de connaissances qui seront rendues accessibles à tous. Sans ce travail d'évaluation, l'information reste simplement « une masse de données distinctes ». C'est ici que l'on constate l'importance du rôle que peuvent jouer les professionnels de l'information dans cette nouvelle société de la connaissance où l'information est abondante, mais aussi où sa qualité et sa diffusion dépendent d'un travail supplémentaire. Dans le même esprit que l'UNESCO, Lytras et Sicilia (2005, p.4), pour leur part, définissent la société de la connaissance comme :

> (...) a new Strategic Position of our Society where the Social and the Economic Perspective is concentrated on the exploitation of emerging technologies, and well-defined knowledge and learning infrastructures are the main vehicles for the implementation of knowledge and learning strategies. The final milestone is a society with access to knowledge and learning for everyone.

Ainsi, c'est par l'utilisation des technologies et des infrastructures de la connaissance qu'il sera possible d'accéder à l'implantation de stratégies d'apprentissage. Les deux auteurs soulignent aussi la nécessité d'accorder une importance particulière au balisage sémantique des ressources d'apprentissage et à la standardisation des métadonnées afin d'assurer l'accessibilité aux connaissances pour tous. Ensuite, lors d'une conférence TED filmée en 2009, Berners-Lee présente le concept du Web de données en appelant les usagers du Web à y partager et à y publier leurs données. Il souligne le fait que les données peuvent prendre des formes différentes et proviennent de tous les domaines, faisant en sorte que ce mouvement peut s'avérer bénéfique à tous. En faisant référence de façon indirecte au concept de la société de la connaissance, il indique que cette nouvelle extension du Web mènera à la création d'informations et de connaissances, mais surtout elle permettra aux machines de créer un sens indépendamment des technologies de l'information et des communications qui, pour leur part, sont plutôt axées sur l'utilisateur. La société de la connaissance se caractérise aussi par le fait qu'on a pu constater, au fil des dernières décennies, un virage au niveau de la production de services. En effet, nous sommes passés d'une consommation de biens physiques à des biens intangibles, des services complètement nouveaux ont vu le jour et l'accessibilité à l'information, avec l'arrivée du Web, s'est vue dépasser les frontières des organisations (Powell et Snellman, 2004).

En tenant compte de cette réalité, il est important que les professionnels du domaine des sciences de l'information se retrouvent au premier plan dans cette nouvelle ère caractérisée par le développement rapide des technologies ainsi que la gestion de l'information. Ainsi, comment les institutions documentaires peuvent-elles mettre en pratique les technologies du Web sémantique et du Web de données dans l'objectif de répondre aux besoins nés de cette société de la connaissance ? Comment le Québec peut-il se démarquer et jouer un rôle au sein de ce changement de paradigme ? Les concepts de « virage numérique » et de « ville intelligente » sont de plus en plus abordés dans les milieux politiques et économiques. On constate une sensibilité de plus en plus importante à l'implantation et à la mise en pratique de nouvelles technologies[1] dans l'objectif non seulement d'améliorer la qualité de vie des citoyens, mais

[1] On n'a qu'à penser au *Plan culturel numérique du Québec* ou encore à *Montréal, ville intelligente et numérique*.

aussi d'assurer une accessibilité aux informations et aux savoirs, une plus grande ouverture et transparence ainsi qu'une prospérité économique.

Comme la tâche est complexe et les enjeux importants, il est opportun d'évaluer les différentes étapes et ressources nécessaires à l'implantation d'un projet de Web de données au sein d'une bibliothèque. Nous avons fait l'exercice en nous penchant sur l'exemple de la Bibliothèque nationale de France (BnF), dont le projet innovateur de Web de données est bien avancé. Pour sa part, Bibliothèque et Archives nationales du Québec (BAnQ) est une institution dont la mission est semblable et qui fait office d'autorité au sein de la francophonie de l'Amérique du Nord. L'objectif est donc d'élaborer des recommandations dans le but de les présenter à cette institution, en vue d'une éventuelle application des technologies du Web sémantique en son sein.

1.3 Méthodologie

Dans le cadre de ce mémoire, nous avons d'abord procédé à une revue de la littérature approfondie sur les technologies qui constituent le Web sémantique et le Web de données. Ensuite, à partir de ce travail, une étude des enjeux et de la pertinence du Web sémantique pour les professionnels de l'information a été effectuée. Puis, nous avons travaillé à une analyse comparative entre les deux institutions (BAnQ et BnF) dans l'objectif d'élaborer une liste présentant les étapes pour la publication de données dans le Web de données. La liste de ces treize étapes a été créée dans le but de synthétiser et de regrouper les informations disponibles, mais disséminées dans la littérature sur le sujet.

Nous avons ensuite procédé à une description des états d'avancement du projet de Web de données au sein de chacune des institutions. Les mises en œuvre de la BnF sont nettement plus avancées que celles de BAnQ, ce pour quoi nous tentons d'évaluer dans quelle mesure il nous serait possible d'apprendre de leurs pratiques pour éventuellement proposer un plan semblable au Québec. Les informations que nous avons recueillies sont d'abord le produit d'une revue de la littérature, mais aussi d'entrevues informelles, par courriels, avec les responsables

du projet data.bnf.fr. Grâce à notre participation à l'École d'été francophone en sciences de l'information à Lyon durant l'été 2015, nous avons pu échanger avec Madame Pauline Moirez, experte des techniques documentaires numériques et des services en ligne à la BnF. Cette dernière nous a permis d'entrer en contact avec Madame Agnès Simon, responsable du projet data.bnf.fr. Nous avons donc pu lui envoyer nos questions par courriel ainsi qu'à Madame Raphaëlle Lapôtre qui a assuré l'intérim d'Agnès Simon à partir de l'hiver 2015.

Du côté de BAnQ, les informations présentées dans ce mémoire sont le résultat d'échanges privilégiés, sous la forme d'entrevues informelles, avec l'équipe chargée de ce type d'initiatives et avec le directeur de l'architecture et de la conception à la Direction générale des technologies de l'information et des télécommunications, Monsieur Jean-François Gauvin. Nous avons pu les rencontrer à deux reprises afin de poser des questions en lien avec ce travail. De plus, par le biais de notre implication au sein du Comité d'expertise sur le Web sémantique mis sur pied dans le cadre de la mesure 06 du Plan culturel numérique du Québec, sur lequel nous reviendrons à la section 5.2, nous avons pu analyser les informations relatives aux projets envisagés du côté de BAnQ.

Finalement, nous avons effectué l'élaboration de recommandations basées sur l'analyse comparative entre les deux institutions et un choix de méthodologie de développement informatique (le *Rational Unified Process* (RUP)) qui semble adapté à la situation décrite. Notons que la liste des étapes présentée ne correspond pas nécessairement dans son entièreté au processus de publication utilisé par la BnF ou encore au processus envisagé par BAnQ. En effet, à chaque institution ses particularités et l'objectif lié à la présentation de ces étapes et de ces pratiques est de proposer une ligne directrice simple pour les institutions.

1.4 Structure du mémoire

Ce mémoire se présente en deux parties. D'abord, une première partie (chapitres 2 et 3) théorique et technique permet de présenter les concepts de base nécessaires à la compréhension de la seconde partie (chapitres 4 à 6). Les sections 2.4 à 2.6 visent à définir de façon plus précise certaines technologies qui seront abordées dans la seconde partie. Cela peut paraître comme une

entrée en matière un peu abrupte, mais la compréhension de celles-ci est nécessaire afin de pouvoir bien saisir les enjeux. C'est pourquoi elles sont présentées dès le second chapitre. Le lecteur peut choisir de différer la lecture de ces sections. Il peut par ailleurs se référer au chapitre 2 en cours de lecture du mémoire pour revenir aux définitions. Un glossaire présentant des définitions sommaires est aussi disponible en annexe 3.

En ce qui a trait à la seconde partie du mémoire, soit les chapitres 4 à 6, elle correspond à l'analyse comparative entre les deux institutions et à la présentation de la liste des étapes nécessaires à la mise sur pied d'un projet de Web de données en bibliothèque. Ainsi, le chapitre 4 présente une description générale de chaque bibliothèque nationale, toujours en priorisant uniquement les aspects pertinents à une comparaison entre celles-ci. Le chapitre 5, pour sa part, propose une description et une comparaison de leurs projets liés au Web sémantique. Finalement, les différentes étapes relatives à l'adoption de ces technologies ainsi que des recommandations seront présentées au chapitre 6.

2. Définition des concepts

Ce chapitre a comme objectif de clarifier certains concepts afin de faciliter la compréhension de la situation actuelle et des occasions d'innovation que le Web sémantique et le Web de données peuvent créer. D'abord, nous présenterons la signification des expressions Web sémantique (section 2.1) et Web de données (section 2.2). Ensuite, nous décrirons le concept de l'Internet des objets (section 2.3) ainsi que celui de données ouvertes (section 2.4). La section 2.5, plus technique, présente le *Semantic Web Stack*, qui permet une visualisation de la hiérarchisation des technologies et standards qui construisent le Web sémantique. Cette illustration est découpée en quatre groupes : identification, description, logique et requêtes. La section 2.6, présente pour sa part les concepts de vocabulaires et d'ontologies et proposent des exemples d'utilisations fréquentes. Finalement, la section 2.7 définit le concept de données structurées internes, une application du Web sémantique souvent mise sur pied parallèlement au Web de données.

2.1 Web sémantique

L'expression « Web sémantique » (*Semantic Web*) est attribuée à Berners-Lee, Hendler et Lassila (2001) qui le décrivent non pas comme un Web distinct de celui que l'on connaît, mais plutôt comme une extension de celui-ci. Il s'agit d'un ensemble de technologies dont l'objectif est d'exploiter de manière plus efficace la quantité importante d'informations que l'on retrouve sur le Web. De grandes quantités de données structurées sont présentement conservées dans des bases de données isolées du Web, que l'on pourrait comparer à des silos d'informations, dont font partie les métadonnées des catalogues de bibliothèques. L'objectif est de les rendre accessibles et de les encoder à l'aide de standards et de normes bien définis, de manière à ce que les machines puissent les interpréter et à assurer une meilleure collaboration entre humains et machines. Ainsi, la compréhension et l'interprétation du sens que prennent les mots et les métadonnées ne se limitent plus à l'humain. Cette interprétation permet, par la suite, de gérer de manière plus efficace (et de manière automatique) l'information, de permettre la

création de nouveaux services et applications ainsi que d'assurer l'accessibilité aux connaissances à la fois par l'humain et par la machine.

Dès le début de l'existence du Web, on évaluait déjà la possibilité d'organiser l'information de façon à ce qu'elle soit « comprise » par la machine (Berners-Lee, Cailliau, Luotonen, Nielsen et Secret, 1994; Shadbolt, Hall et Berners-Lee, 2006). Le Web est un espace virtuel où l'information est disponible dans une grande variété de langues naturelles et dont le public est humain tandis que, dans le Web sémantique, l'information est exprimée dans un langage destiné à être compris et interprété par des machines (DeWeese et Segal, 2015). Le Web sémantique est donc une infrastructure qui fournit un sens aux données et aux documents grâce à l'ajout de relations entre les métadonnées qui les caractérisent. Soulignons le fait que le Web sémantique est basé sur des concepts déjà bien établis et assimilés, tels que le protocole *Hypertext Transfer Protocol* (HTTP) ainsi que les *Uniform Resource Identifiers* (URI). À la section 2.5, nous décrirons plus en détail les différentes composantes technologiques qui en permettent le développement.

2.2 Web de données

Le Web de données (*Linked Data*), pour sa part, est l'une des composantes du Web sémantique et il s'agit probablement de l'application la plus connue. Il arrive d'ailleurs parfois que certains confondent les deux et plusieurs ne s'entendent pas encore sur ce point (Heath, 2009; Hendler, 2009). D'ailleurs, depuis le 11 décembre 2013, le W3C a cessé de mettre à jour la page Web de son site dédié au Web sémantique[2] en indiquant aux visiteurs que celle-ci avait été absorbée par une autre page se consacrant à ce qu'ils nomment le *W3C Data Activity*[3]. Il est donc intéressant de souligner l'évolution du concept du Web sémantique, non seulement aux yeux des professionnels du domaine, mais aussi au sein même de l'organisation qui le chapeaute. Le W3C indique d'ailleurs que les résultats de l'implantation du Web sémantique démontrent

[2] http://www.w3.org/2001/sw

[3] http://www.w3.org/2013/data

un certain succès, mais qu'ils ne correspondaient pas à ce qui avait été envisagé au départ. En effet, le consortium s'attendait à ce que les activités de publication de pages Web et de données soient similaires. Cependant, on a rapidement constaté que la publication de données (dates, titres, propriétés chimiques, lieux, etc.) est vue comme une activité effectuée par des spécialistes du domaine et non pas comme une activité que tous pourraient accomplir (W3C, 2013).

Afin d'éviter toute confusion entre Web sémantique et Web de données, il est possible de dire que le Web de données est l'une des applications qui entre dans la grande famille du Web sémantique. Comme nous allons le voir à la section 2.7, les données structurées internes sont aussi une des applications du Web sémantique. Le Web de données doit être considéré comme l'ensemble des pratiques et des standards permettant de publier des données structurées sur le Web afin que celles-ci puissent être liées entre elles et interrogées.

En 2006, Tim Berners-Lee publie le premier jet d'un article sur le site du W3C intitulé *Linked Data*. Cette page a été modifiée plusieurs fois jusqu'en 2010 avec l'ajout du concept de *Linked Open Data 5 Star* sur lequel nous reviendrons sous peu. Cette note a comme objectif de clarifier certains aspects du Web de données et de rappeler les objectifs derrière l'implantation du projet, soit un partage des données et le fait d'effectuer des liens entre elles. La mise sur pied du groupe d'intérêt *Semantic Web Education and Outreach Interest Group* (SWEO), maintenant fermé depuis 2008, a donc permis d'assurer l'accessibilité à des guides permettant à ceux qui le souhaitaient de mieux comprendre les outils qui étaient mis à leur disposition pour pouvoir publier leurs données liées sur le Web.

Il s'agissait aussi de promouvoir l'idée selon laquelle les données étaient comparables aux documents et qu'il était possible de les relier entre elles de la même façon que les documents sont liés entre eux grâce aux liens hypertextes sur le Web (Bermès, Isaac et Poupeau, 2013; Bizer et Heath, 2011). L'initiative *Linking Open Data community project*[4] mise sur pied par le SWEO a permis de sensibiliser les communautés qui encodaient déjà leurs données pour le Web

[4] http://linkeddata.org

sémantique à les publier sous licence libre et à les relier entre elles. Ce projet a donné naissance à DBpedia[5], un projet communautaire piloté par l'Université de Leipzig, l'Université de Mannheim et l'entreprise OpenLink Software. Il s'agit d'une base de données composée d'informations encodées dans le format standard du Web sémantique extraites de Wikipédia. DBpedia a permis aux usagers d'avoir accès à un assez grand nombre de données provenant de plusieurs domaines différents pour pouvoir relier leurs propres données à un point d'ancrage (Bermès, Isaac et Poupeau, 2013). En mars 2012, une version francophone[6] a été présentée par l'Institut national de recherche en informatique et en automatique (Inria), le Ministère de la Culture de la France ainsi que l'association Wikimédia France.

Le Web de données est souvent représenté par un graphe, attribué à Max Schmachtenberg, Christian Bizer, Anja Jentzsch et Richard Cyganiak, qui permet de constater les différents liens entre les nombreux jeux de données (des collections de données) (voir annexe 1). Les jeux de données qui se trouvent dans le graphe répondent aux critères suivants (Schmachtenberg, Bizer, Jentzsch et Cyganiak, 2014) :

- Les URI doivent être déréférençables (nous y reviendrons à la section 2.5.1.1) ;
- Les données doivent être des données RDF et être encodées dans l'une des syntaxes de sérialisation les plus utilisées (RDF/XML, Turtle, N-Triples) (nous y reviendrons à la section 2.5.2) ;
- Le jeu de données doit contenir au moins 1 000 triplets ;
- Le jeu de données doit être lié à au moins un autre jeu de données se trouvant dans le graphe ;
- L'accès au jeu de données dans son entièreté doit être possible grâce des *dumps* ou un point d'accès SPARQL (nous y reviendrons à la section 2.5.4).

Les jeux de données les plus importants et vers lesquels le plus grand nombre de liens ont été effectués sont représentés par des sphères plus larges dans le graphe. Ainsi, les DBpedia[7],

[5] http://wiki.dbpedia.org

[6] http://fr.dbpedia.org

[7] http://wiki.dbpedia.org/

GeoNames[8] (une base de données géographique) et FOAF[9] (un vocabulaire qui permet de décrire les personnes et les relations qui les unissent) se retrouvent au centre et sont plus imposants que les autres, ce qui démontre leur importance et leur poids dans le Web de données. Ce graphe est en constante évolution et il est intéressant de comparer les différentes versions disponibles afin de bien constater à quel point le mouvement est en croissance (voir figure 1).

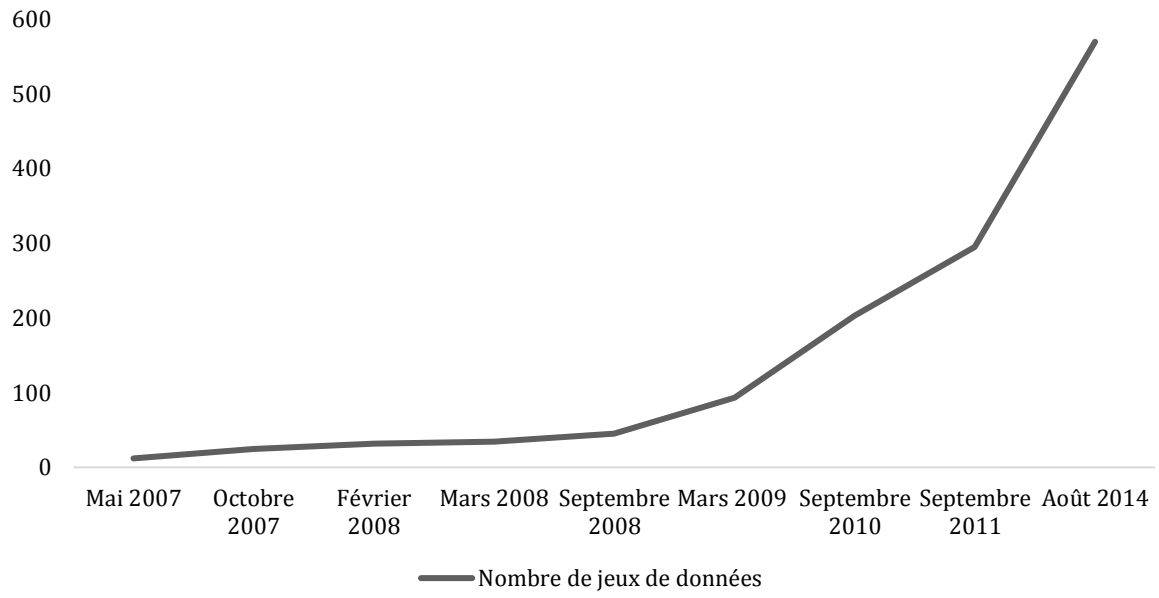


Figure 1. Évolution du nombre de jeux de données publiés selon les standards du Web de données et interreliés à d'autres jeux de données sur le Web, de mai 2007 à août 2014[10]

Dans sa note de 2006, Berners-Lee présente les quatre principes de base du Web de données sur lesquels plusieurs se baseront par la suite et sur lesquels nous reviendrons :

1. Nommer les ressources avec des URI ;
2. Utiliser des URI déréférençables (protocole HTTP) afin qu'il soit possible d'accéder à des informations sur les ressources ;

[8] http://www.geonames.org/ontology/documentation.html

[9] http://www.foaf-project.org/

[10] Source : http://lod-cloud.net/

3. S'assurer que les URI déréférençables fournissent des informations pertinentes à l'aide des standards tels que RDF et SPARQL ;
4. Créer un réseau de liens avec d'autres URI provenant d'autres bases de données.

2.3 Internet des objets (IdO) et Web des objets

Il est important, afin d'éviter toute confusion, de définir aussi les concepts de l'Internet des objets (*Internet of Things* (IoT)) et du Web des objets (*Web of Things* (WoT)). D'abord, l'entreprise américaine Gartner qui se spécialise dans le conseil et la recherche dans le domaine des technologies de l'information définit l'Internet des objets comme : « (…) the network of physical objects that contain embedded technology to communicate and sense or interact with their internal states or the external environment. » (Gartner, 2016). Plusieurs définitions existent et celles-ci évoluent en même temps que les technologies, mais il semble y avoir consensus pour définir l'Internet des objets comme une infrastructure globale pour la société de la connaissance qui permet la création de services avancés grâce à l'interconnexion d'objets virtuels et réels (souvent qualifiés de *smart*), basée sur les technologies de l'information. Ces technologies sont interopérables et évolutives (ITU, 2016; Kopetz, 2011; Mulani et Pingle, 2016).

Le Web des objets, pour sa part, est une couche applicative visant à réutiliser des standards du Web déjà maîtrisés afin de faciliter la création d'applications dans l'Internet des objets (Guinard, Trifa, Mattern et Wilde, 2011). Il est important de ne pas confondre le Web des objets ou l'Internet des objets avec les technologies du Web sémantique décrites dans le cadre de ce travail, bien que celles-ci pourraient éventuellement être considérées comme des solutions pour la réalisation de l'IdO (Barnaghi, Wang, Henson et Taylor, 2012; Zaino, 2013).

2.4 Données ouvertes

Parallèlement au mouvement du Web de données, on a constaté au cours des dernières années un intérêt marqué pour le mouvement des données ouvertes, encourageant l'accès à l'information, à la réutilisation et à la redistribution de données. Le Web a joué un rôle important dans ce mouvement dans la mesure où il facilite l'accès et le partage tout en fournissant un lieu

neutre d'échange, de publication et de stockage d'information numérique peu coûteux. Ainsi, dans le domaine de la science, on constate l'existence de la notion de science ouverte (*open science*) dont l'objectif est de collaborer, de produire des hypothèses, de présenter des méthodologies, de rendre disponibles les résultats et les données de recherches scientifiques et d'arriver à un résultat qui aura permis la contribution de tous ceux qui auraient l'expertise nécessaire. Ensuite, le mouvement des données ouvertes a particulièrement été mis en place au sein de gouvernements qui ont pris l'initiative de rendre leurs données accessibles dans l'optique d'améliorer la transparence, d'augmenter la participation citoyenne et de faciliter la collaboration. Les gouvernements du Canada[11] et du Québec[12] ont d'ailleurs tous les deux mis sur pied des plateformes donnant accès à des jeux de données ouvertes. Soulignons aussi la présence de l'*Open Knowledge Foundation*[13], une association à but non lucratif promouvant l'ouverture des données ainsi que la culture et les contenus libres.

Dans le cadre du Web de données, le mouvement de données ouvertes est intéressant, car on constate la possibilité de jumeler les deux principes et ainsi non seulement rendre accessibles des données, mais aussi les interrelier. Dans sa note sur les données liées (2006), Berners-Lee a, en 2010, ajouté une échelle de qualité basée sur cinq étoiles qui permet d'évaluer jusqu'à quel point l'information est facilement réutilisable. Les cinq étapes (ou étoiles) sont les suivantes et la figure 2 permet de les illustrer :

1. Rendre vos données disponibles sur le Web (quel que soit leur format) en utilisant une licence ouverte (*Open Licence*) ;
2. Rendre vos données disponibles sous forme de données structurées (par exemple, en format Excel plutôt que sous forme d'image numérisée d'un tableau) (*Reusable*) ;
3. Utiliser des formats non-propriétaires (par exemple, Comma-separated values (CSV) plutôt qu'Excel) (*Open Format*) ;

[11] http://ouvert.canada.ca/fr

[12] http://www.donnees.gouv.qc.ca

[13] http://okfn.org

4. Utiliser des URI pour identifier vos données afin que les autres utilisateurs puissent pointer vers elles (*URI*) ;
5. Relier vos données à d'autres données pour fournir du contexte (*Linked Data*).

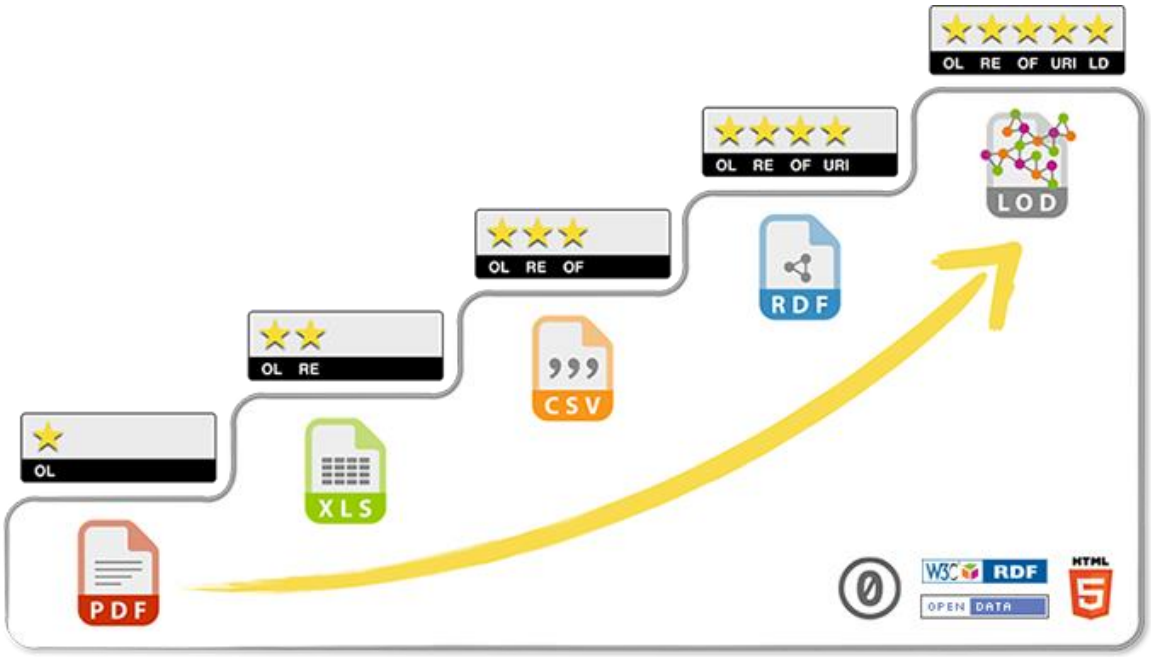


Figure 2. Échelle de qualité des données ouvertes proposée par Berners-Lee (2010)[14]

Pour les institutions culturelles telles que les bibliothèques, les données ouvertes et liées (DOL) sont donc une application du Web de données visant à ouvrir les données à la réutilisation. Il est possible de mettre sur pied des applications du Web de données à l'interne, mais cette façon de faire n'est pas recommandée pour les institutions culturelles dans la mesure où elle ne va pas de pair avec leurs missions. Ainsi, en général, on parlera de Web de données en impliquant inévitablement une ouverture de ces données aux usagers et aux développeurs Web.

[14] Source : http://5stardata.info/en/

2.5 *Semantic Web Stack*

Cette section commence la description plus technique de certaines technologies nécessaires à l'implantation d'un projet de Web de données et c'est à partir d'ici que la mise en garde présentée à la section 1.4 s'applique. Nous rappelons au lecteur qu'il peut se référer aux sections à venir à tout moment au cours de sa lecture ou au glossaire qui se trouve en annexe 3.

L'architecture du Web sémantique est souvent représentée à l'aide d'une illustration nommée *Semantic Web Stack*, *Semantic Web Cake* ou encore *Semantic Web Layer Cake* (voir figure 3). En français, on retient les expressions Pyramide de standardisation ou encore Pyramide du Web sémantique. Cette illustration permet de décrire la hiérarchisation des technologies informatiques, représentées par les couches de la pyramide, qui constituent le Web sémantique. Ces technologies sont en évolution, dans la mesure où certaines d'entre elles sont toujours en voie de normalisation et celles qui composent la partie supérieure de la pyramide et la façon dont elles seront implantées sont encore à l'étude.

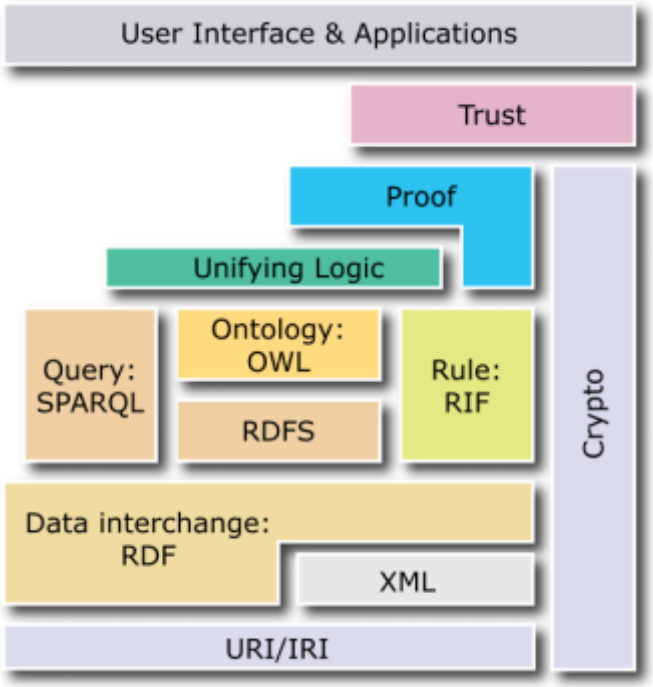


Figure 3. Le *Semantic Web Stack* proposé par le W3C (2007)

En analysant cette pyramide, il est possible de la découper en six groupes différents :

1. Identification : URI, IRI ;
2. Description : XML, RDF ;
3. Logique : OWL, RDFS, RIF, *Unifying Logic* ;
4. Requêtes : SPARQL ;
5. Confiance et sécurité : *Crypto*, *Trust*, *Proof* ;
6. Interaction avec l'usager : *User interface*.

Notons que ce découpage a comme but de décrire de manière générale les technologies et quelques-unes d'entre elles peuvent participer à plus d'un objectif. D'abord, afin de bien comprendre ces différents groupes, il est nécessaire de revenir à l'article de Berners-Lee, Hendler et Lassila paru dans la revue *Scientific American* en 2001. Dans ce texte, les auteurs présentent une vision particulière de ce que pourra impliquer concrètement l'arrivée du Web sémantique. Dans le cadre des sciences de l'information, on pourrait donner l'exemple suivant : grâce aux technologies du Web sémantique, il serait possible d'identifier de manière automatique des auteurs qui partageraient un certain nombre de caractéristiques (même origine linguistique ou géolinguistique, par exemple), sur la base d'informations puisées à même différents jeux de données de provenances différentes.

Le système est basé sur la présence d'agents logiciels indépendants et autonomes caractérisés par une intelligence artificielle forte[15]. Une fois la requête faite, ces agents parcourent les jeux de données pour obtenir la réponse la plus satisfaisante pour l'usager. Le travail préalable d'encodage et de construction d'ontologies (voir la définition présentée à la section 2.6 ou le glossaire en annexe 3) dépend toujours de la présence d'un humain, mais le fonctionnement réside dans le fait que la logique acquise par les agents peut ensuite être réutilisée. Les interprétations de ce qu'est et de ce que pourrait être le Web sémantique sont nombreuses et cet article accorde une importance particulière à ce concept d'agents, mais aussi aux concepts d'échange de preuves (*proof*), de langage unifié (*unifying language*) et de

[15] Selon l'entrée de Copeland (2016) dans l'*Encyclopædia Britannica*, l'intelligence artificielle forte a comme objectif de construire des machines qui pensent et dont les capacités intellectuelles sont égales à celles d'un être humain.

signatures numériques (*digital signatures*). Ces concepts composent les groupes cinq et six présentés précédemment, soit la confiance et la sécurité ainsi que l'interaction avec l'usager. Comme mentionné, ces groupes sont encore à l'étude et l'on ne peut, pour le moment, décrire leur fonctionnement en profondeur. Cependant, on peut comprendre que la couche *proof* fait référence à la provenance de l'information générée par l'agent ainsi qu'à sa fiabilité (Hart et Dolbear, 2013; Schillinger, 2011). Par exemple, les notices d'autorité indiquent un certain niveau de fiabilité quant aux résultats générés par les agents logiciels. Directement liée à cette couche, on retrouve la logique unificatrice (*unifying logic*), qui permet de traduire le raisonnement des agents et, ainsi, permettre d'avoir accès à la preuve que l'information donnée est fiable.

Dans le cadre du Web de données, il a été démontré qu'il est possible de créer de nouveaux services et des applications uniquement à l'aide des technologies recommandées et éprouvées par le W3C, donc les groupes un à quatre présentés ci-dessus (identification, description, logique et requêtes) (Baker et al., 2011; Bizer et Heath, 2011). Dans les sections suivantes, nous nous pencherons sur ces groupes du *Semantic Web Stack* afin de mieux comprendre les différents standards qui composent cette nouvelle extension du Web et leurs fonctionnalités.

2.5.1 Identification

Cette section présente les concepts d'URI et d'IRI, nécessaires à la compréhension des sections 3.2.7 et 6.1.7 de ce mémoire.

Dans le Web sémantique, les sujets, les objets ainsi que les « verbes » qui les relient doivent être représentés par des URI ou des *Internationalized Resource Identifier* (IRI), qui sont des identifiants, c'est-à-dire des suites alphanumériques qui identifient de manière univoque des ressources physiques ou abstraites. La différence entre les URI et les IRI réside dans le fait que les URI sont limités à l'utilisation du jeu de caractères *American Standard Code for Information Interchange* (ASCII), qui permet d'écrire en anglais, alors que les IRI peuvent être construits à partir du *Universal Coded Character Set (UCS* – norme ISO/IEC 10646), qui permet pour sa

part l'utilisation de caractères provenant d'autres langues telles que le chinois, le japonais, le coréen, etc. Par exemple, on pourrait représenter l'auteur Yukio Mishima avec l'URI suivant :

http://www.exemple.com/auteurs/yukio_mishima

Il ne serait alors pas possible de le représenter à l'aide des caractères japonais. Cependant, les IRI permettent cette construction, donc il serait possible de représenter l'auteur ainsi, si nécessaire :

http://www.exemple.com/auteurs/三島由紀夫

Un autre exemple pourrait être celui de l'utilisation des accents en langue française. Les URI ne permettent pas l'utilisation des accents, donc si l'on souhaitait proposer l'URI :

http://www.exemple.com/auteurs#honoré_de_balzac

Celui-ci ne serait point fonctionnel et apparaîtrait ainsi :

http://www.exemple.com/auteurs#honor%c3%a9_de_balzac

Pour sa part, l'IRI permettrait alors l'utilisation de l'accent aigu. De plus en plus, on remarque que l'utilisation d'IRI est préconisée, car elle contribue à une plus grande interopérabilité et à l'universalité du Web. Cependant, pour le moment, on constate une utilisation plus importante des URI au sein de la communauté alors nous allons nous y attarder plus longtemps afin d'être conséquents avec les choix effectués par la majorité. Étant donné que les IRI n'ont été standardisés qu'en 2005 et que les institutions qui ont mis sur pied des projets de Web sémantique n'avaient pas à recourir à un jeu de caractères autre que l'ASCII, l'usage des URI est plus fréquent (Gangemi et Presutti, 2006).

Les URI sont des identifiants dont la syntaxe est normalisée par le W3C qui permettent d'identifier des ressources sur un réseau donné. La syntaxe (ou la façon de les exprimer) est la suivante (Bermès, Isaac et Poupeau, 2013) : schème:autorité/chaîne_de_caractères. Dans le cas qui nous intéresse, le schème utilisé pour construire les URI est le schème http: . La grande majorité des utilisateurs du Web sont déjà familiers avec les URI étant donné l'utilisation fréquente des *Uniform Resource Locator* (URL), un type d'URI, qui permettent l'identification ainsi que l'accès aux ressources sur le Web (p. ex., http://www.umontreal.ca/). Notons aussi l'existence des *Uniform Resource Name* (URN), qui sont aussi un type d'URI, qui permettent

pour leur part d'identifier une ressource durant toute la durée de son existence, et ce, indépendamment de sa location ou de son accessibilité (voir figure 4).

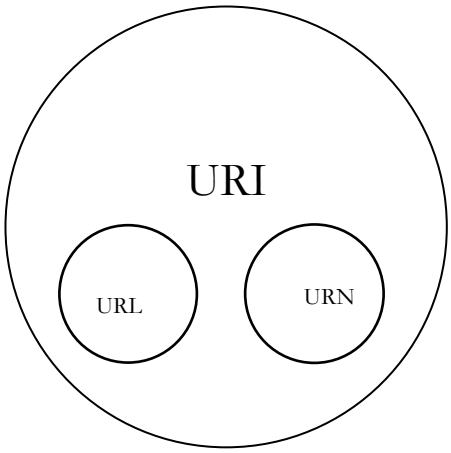


Figure 4. Diagramme de Venn illustrant les liens unissant URI, URL et URN

Les URI permettent d'identifier ou de nommer une ressource, qui peut être n'importe quels objet ou concept existant dans le monde. Berners-Lee, Fielding et Masinter (2005) donnent d'ailleurs des exemples de ce que l'on peut considérer comme des ressources, soit un document numérique, une image, un service. Ils soulignent aussi le fait que les ressources n'ont pas l'obligation d'être accessibles via Internet et qu'ainsi, un être humain, un organisme, un livre emprunté à la bibliothèque ou même un type de relation peuvent tous être considérés comme des ressources. Vatant (2008) indique donc qu'une ressource peut être entièrement matérielle, mais aussi strictement virtuelle et dématérialisée. Ainsi, un URI peut identifier une ressource existante dans le monde réel comme, par exemple, Réjean Ducharme, mais un autre URI peut identifier aussi le concept de Réjean Ducharme comme sujet d'un document donné.

2.5.1.1 URI déréférençables et négociation de contenu

Cette section présente des spécificités techniques plus approfondies cependant nécessaires à la compréhension de la section 6.1.7, principalement. Comme indiqué précédemment, le second principe du Web de données est « utiliser des URI déréférençables

(protocole HTTP) afin qu'il soit possible d'accéder à des informations sur les ressources »[16]. Afin de comprendre ce qu'est un URI déréférençable, il est nécessaire de comprendre le concept de négociation de contenu proposé par Berners-Lee (2006). Lorsqu'un usager ou une machine parcourt les liens de manière intuitive (modèle d'interopérabilité par liens ou *follow your nose* (Bermès, Isaac et Poupeau, 2013)), il est nécessaire que ceux-ci puissent obtenir de l'information qui leur sera intelligible et qu'ils pourront traiter. Ainsi, un URI doit présenter une page HTML si le demandeur est un humain et une sérialisation de RDF (nous y reviendrons à la section 2.5.2) si le demandeur est une machine. La requête HTTP sur une ressource informera le serveur de la nature de l'usager : un navigateur Web signifie qu'il s'agit d'un humain et un programme informatique ou une application du Web sémantique signifie qu'il s'agit d'une machine (Bermès, Isaac et Poupeau, 2013; Zengenene, Casarosa, Meghini, 2014). Ainsi, lorsqu'un URI identifie un objet existant dans le monde réel, il est nécessaire d'assurer une distinction entre cet objet, le document Web qui le décrit et le flux RDF destiné à la machine. Pour une même ressource, on créera donc en général au moins trois URI différents (Berners-Lee, 2008; Zengenene, Casarosa et Mehini, 2014) :

- un URI pour décrire l'objet réel (ou le concept);
    - http://www.exemple.com/ark:/12345678
- un URI pour décrire la représentation HTML;
    - http://www.exemple.com/12345678/gabrielle_roy
- un URI pour décrire la représentation RDF/XML.
    - http://www.exemple.com/12345678/gabrielle_roy/rdf.xml

2.5.1.2 Définition univoque des relations entre les URI

Ces trois URI doivent par la suite être liés entre eux afin que leur relation soit bien définie. Ainsi, un exemple de mécanisme de redirection mis en place à partir des URI pourrait être : http://www.exemple.com/ark:/12345678 (URI du concept) redirige par HTTP 303 vers

[16] Notons qu'un URI peut aussi permettre d'accéder à une ressource Web et non uniquement à des informations sur celle-ci.

http://www.exemple.com/12345678/gabrielle_roy (adresse de la page HTML). HTTP 303 (*see other*) est un code visant à rediriger les applications Web vers un nouvel URI. Ainsi, dans la mesure où la requête provient d'un navigateur Web, cela signifie que le demandeur est humain, donc qu'on doit le rediriger vers l'URI de la page HTML. La figure 5, basée sur la note du W3C de 2008 et sur l'exemple ci-dessus, illustre comment les trois URI doivent être liés les uns aux autres :

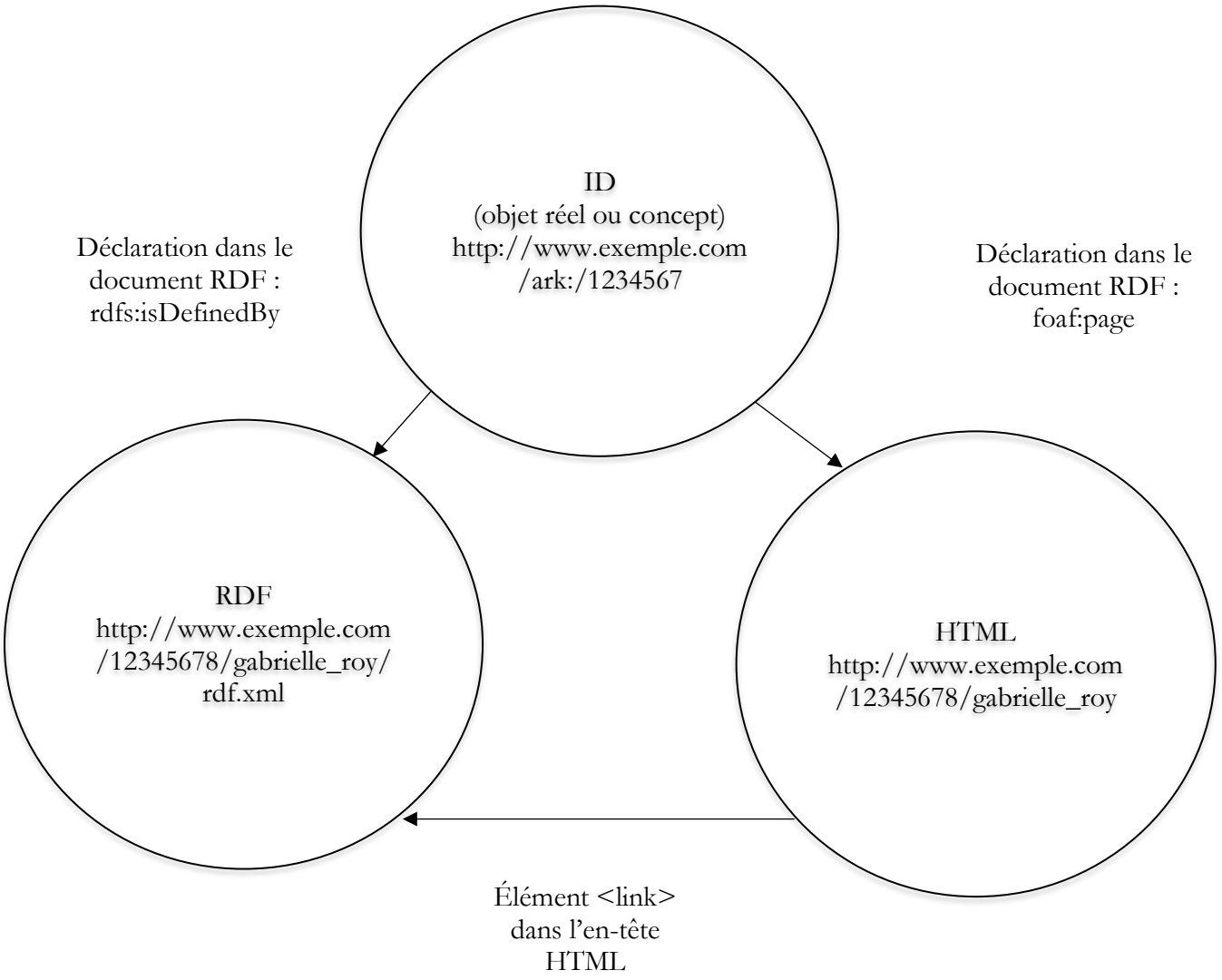


Figure 5. Relation entre trois URI (objet réel, RDF et HTML) décrivant la même ressource

Ainsi, le document RDF que l'on retrouve grâce à l'identifiant http://www.exemple.com/12345678/gabrielle_roy/rdf.xml doit contenir les déclarations visant à assurer les liens avec les deux autres identifiants. L'utilisation de la propriété provenant de RDF Schema rdfs:isDefinedBy (nous y reviendrons à la section 2.5.3) permet d'indiquer que l'URI identifiant l'objet réel est le sujet du document RDF. Ensuite, l'utilisation de la propriété provenant de *Friend of a friend* (nous y reviendrons aussi à la section 2.5.3) permet d'indiquer que le document HTML identifié par l'URI http://www.exemple.com/12345678/gabrielle_roy a comme sujet l'URI identifiant l'objet réel. Finalement, il est nécessaire de créer un lien entre le document RDF et le document HTML. Pour ce faire, on utilise l'élément d'en-tête <link> qui permet de spécifier un lien vers un autre document.

Les URI déréférençables ont donc l'avantage de permettre la récupération d'une ressource, mais aussi de fournir des informations au sujet de ressources qui ne peuvent être récupérées via le Web (Bizer, Heath et Berners-Lee, 2009). Aussi, plus les options visant à naviguer d'une donnée à l'autre et de télécharger les données sont nombreuses, plus on a de chance d'obtenir une meilleure visibilité et un plus grand achalandage. L'utilisation d'URI pour identifier les ressources permet de diminuer l'ambigüité possible, non seulement en ce qui a trait aux ressources identifiées, mais aussi au lien qui les unit. Par exemple, il est possible d'utiliser le schéma de métadonnées normalisé Dublin Core[17] pour décrire les relations entre les ressources. Un schéma de métadonnées vise à définir un ensemble fini et formellement défini de métadonnées utilisées pour décrire des ressources. Il permet que les données soient par la suite interprétées de la même manière par tous. Pour s'assurer d'atteindre cette compréhension générale, il est nécessaire de définir les propriétés et caractéristiques des données. Au fil du temps, plusieurs schémas de métadonnées ont été développés pour différents domaines d'étude, dont certains ont été normalisés.

Revenons maintenant à notre exemple faisant appel au schéma de métadonnées Dublin Core. Notons qu'il serait possible d'utiliser de façon directe cet URI

[17] http://dublincore.org/

http://dublincore.org/2010/10/11/dcelements.rdf#creator pour décrire la relation entre un auteur et une œuvre et, ainsi, éviter toute confusion possible (Stuart, 2011). Cela permet de préciser que la notion de « créateur » (*creator*) utilisée est celle qui est définie dans le schéma de métadonnées Dublin Core (celle utilisée notamment pour encoder l'auteur de documents), et non pas une autre notion de créateur qu'on pourrait retrouver ailleurs. Dublin Core propose aussi un élément « contributeur » (*contributor*) qui est défini différemment que l'élément « créateur ». Cette utilisation permet donc une meilleure précision au niveau de la description des ressources. On s'assure ainsi que l'usager aura accès aux bonnes informations sur la ressource et l'on évite l'incertitude par rapport à un concept.

2.5.1.3 Pérennité et accessibilité des URI

Un autre aspect primordial en lien avec la construction d'URI dans le Web de données est d'en assurer la pérennité et l'accessibilité. Bien qu'on encourage déjà fortement cette pratique dans le Web (Berners-Lee, 1998), cet aspect prend une importance encore plus grande dans le cas du Web de données dans la mesure où plusieurs acteurs se basent sur les mêmes URI afin de les relier entre eux. C'est pourquoi la plupart utilisent le site Internet http://purl.org qui permet la création d'URL pérennes et stables. Plusieurs autres exemples de système d'identifiants existent tels que *Archival Resource Key* (ARK) ou encore *Digital Object Identifier* (DOI). Bien que l'objectif principal soit que les URI soient lisibles par les machines, il est aussi nécessaire d'envisager la nécessité de les construire de façon à ce qu'ils soient compréhensibles par l'humain, principalement parce que la programmation se fait encore par des individus (Stuart, 2011). Ainsi, l'URI http://www.exemple.com/auteur/janedoe (identifiant signifiant) sera plus facile à comprendre et réutiliser que http://www.exemple.com/1/190349342 (identifiant opaque).

2.5.2 Description

Cette section présente le modèle RDF de description de données et des relations entre elles, nécessaires à la compréhension des sections 2.6, 2.7 et 6.1.11 de ce mémoire.

Afin d'assurer l'interopérabilité des données, l'organisation de l'information ainsi que la description des ressources et des métadonnées qui les caractérisent, le W3C a développé et normalisé le modèle *Resource Description Framework* (RDF), le langage de base du Web sémantique. RDF est en fait au Web de données et aux Web sémantique ce que HTML est au Web (Bermès, Isaac et Poupeau, 2013). Cette norme est basée sur la formation de triplets constitués d'un sujet, d'un prédicat et d'un objet, tous identifiés à l'aide d'URI pérennes ou d'IRI. Notons que l'objet peut parfois être décrit de façon littérale dans certains cas, ce qui peut cependant mener à une certaine ambigüité et diminuer les chances de création de liens entre les ressources. Ces triplets sont des « phrases » qui permettent donc d'identifier une première ressource (le sujet), l'aspect qui caractérise la relation entre le sujet et l'objet (le prédicat) et la seconde ressource (l'objet) (voir figure 6).

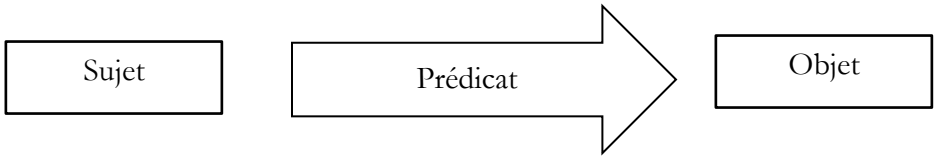


Figure 6. Modèle d'un triplet RDF

Ces triplets peuvent être visualisés en graphes, ce qui permet parfois de mieux saisir le concept. Le tableau 1 présente un exemple d'un graphe sur une ressource donnée.

Tableau 1. Exemple d'une représentation tabulaire d'un graphe RDF

Sujet	Prédicat	Objet
http://www.exemple.com/ressource1	rdf:type	http://schema.org/Book
http://www.exemple.com/ressource1	dc:title	Les enfantômes
http://www.exemple.com/ressource1	dc:creator	http://viaf.org/viaf/115106377 http://lccn.loc.gov/n50026779 http://www.idref.fr/026840588
http://www.exemple.com/ressource1	dc:date	1976
http://www.exemple.com/ressource1	dc:identifier	9782070293537

Ainsi, on constate que ces triplets RDF donnent les informations suivantes grâce, entre autres, au schéma de métadonnées Dublin Core : la ressource 1 (dont l'URI est http://www.exemple.com/ressource1) est un livre dont le titre est *Les enfantômes* et l'ISBN est le 9782070293537, écrit par Réjean Ducharme en 1976. On remarque que le nom de Réjean Ducharme n'est pas exprimé de façon directe dans le tableau. En effet, il est plutôt exprimé à l'aide du prédicat dc:creator qui relie le sujet à d'autres ressources d'autorité correspondant à Réjean Ducharme de la *Library of Congress* (LC), de *Virtual International Authority File* (VIAF) et d'Identifiants et Référentiels (IdRef). Ces trois instances ont pris la décision d'ouvrir leurs fichiers d'autorité afin de permettre la réutilisation et la création de liens vers leurs ressources dans le Web de données.

Afin que ce graphe soit lisible par la machine, on peut ensuite l'encoder à l'aide de ce qu'on appelle « une syntaxe de sérialisation ». Il existe plusieurs syntaxes de sérialisation et formats disponibles, soit RDF/XML, N3, N-Triples, Turtle, JSON, etc. Nous ne présenterons pas les différentes syntaxes en détail ici étant donné leur complexité relative, mais notons qu'en général l'encodage se fait à l'aide d'outils informatiques et non manuellement. Notons aussi que l'annexe 2 propose un exemple simple de triplets RDF représentés en tableau et en graphe. Voici

tout de même un exemple de l'encodage du graphe ci-dessus, dans la syntaxe de sérialisation Turtle[18] (W3C, 2014-a) :

```
@prefix schema: <http://schema.org> .
@prefix dc: <http://purl.org/dc/elements/1.1/> .

<http://www.exemple.com/ressource1>
        a schema:Book;
        dc:title "Les enfantômes" ;
        dc:creator <http://viaf.org/viaf/115106377>,
        <http://lccn.loc.gov/n50026779>, <http://www.idref.fr/026840588> ;
        dc:date "1976" ;
        dc:identifier "9782070293537" .
```

La syntaxe de sérialisation Turtle est la plus facile à lire pour l'être humain. Comme on peut le voir dans l'exemple ci-dessus, Turtle permet l'utilisation d'espaces de nom et de préfixes. Un espace de nom est un lieu virtuel où l'on retrouve l'organisation et la définition de termes particuliers. On peut donc avoir un espace de nom pour une ontologie, un vocabulaire ou un schéma de métadonnées. En les indiquant au début de la description, il est par la suite possible d'y faire référence à l'aide de préfixes. Ainsi, dans l'exemple présenté ci-dessus, les espaces de noms sont "http://schema.org/" ainsi que "http://purl.org/dc/elements/1.1/" et les préfixes sont schema et dc. On utilise donc, dans cette description, le schéma Schema.org ainsi que le schéma de métadonnées Dublin Core. L'utilisation de préfixes permet d'alléger la description. Ainsi, en inscrivant dc:title, on fait référence à l'élément *title* de Dublin Core, sans avoir à indiquer l'URI complet, c'est-à-dire http://purl.org/dc/elements/1.1/title. On constate donc que la syntaxe de sérialisation présente les mêmes informations qui se trouvent dans le tableau 1, mais de façon à ce qu'elles puissent être lues par la machine.

[18] https://www.w3.org/TR/turtle/

2.5.3 Logique

Cette section présente RDFS et OWL, des langages de représentation des connaissances, nécessaires à la compréhension des sections 2.6, 6.1.8 et 6.1.10 de ce mémoire.

RDF Schema (RDFS) et *Web Ontology Language* (OWL) sont des langages de représentation des connaissances recommandés par le W3C ayant comme objectif de fournir une base pour la création d'ontologies qui visent à décrire les entités de notre monde ainsi que leurs relations. Il ne s'agit pas d'ontologies en tant que telles, mais bien des outils qui permettent d'apporter des spécifications à celles-ci. Les ontologies sont des collections de classes et de propriétés (Bermès, Isaac et Poupeau, 2013; Bizer, Heath et Berners-Lee, 2009; Synak, Dabrowski et Kruk, 2009; Vatant, 2008). Nous y reviendrons plus en détail ci-dessous. D'abord, les classes permettent d'obtenir de l'information sur la nature des ressources. Les ressources qui font partie d'une classe donnée sont nommées « instances » de cette classe et les classes elles-mêmes sont des ressources. Ainsi, si la ressource est http://www.example.com/auteur/janedoe/, il s'agit d'une instance de la classe « personne » qui sera déclarée par une ontologie comme, par exemple, *Friend of a friend* (FOAF) qui permet de décrire des personnes ou des organismes. RDFS permettrait donc ici de déclarer la ressource comme une classe pour d'autres ressources. Ainsi, on indiquerait :

http://www.example.com/auteur/janedoe/ rdf:type foaf:Person

Ensuite, les propriétés donnent de l'information sur le prédicat du triplet, donc sur la relation qui unit deux ressources ou une ressource et un littéral. Dans l'exemple ci-dessus, le prédicat rdf:type est une propriété qui indique que le sujet est une instance de la classe indiquée dans l'objet. Notons que les classes et propriétés peuvent aussi être divisées en sous-classes et en sous-propriétés et qu'il y a donc une organisation hiérarchique qui permet une meilleure précision. Par exemple, la classe « lieu » pourrait être ensuite divisée et on obtiendrait la sous-classe « ville » (Bermès, Isaac et Poupeau, 2013). RDFS permet donc de décrire de façon simple, mais limitée, cette logique hiérarchique entre les classes, les sous-classes, les propriétés et les sous-propriétés. OWL vient jouer le même rôle que RDFS, mais en ajoutant plus de valeur sémantique au schéma, donc plus de précision. Par exemple, la propriété owl:sameAs, très utilisée, permettra de lier deux URI et indiquera à la machine qu'il s'agit du même concept, bien

que les données soient dans des jeux de données différents. OWL permet aussi de clarifier la logique hiérarchique afin de faciliter la compréhension par la machine.

2.5.4 Requêtes

Cette section présente le langage de requête SPARQL nécessaire à la compréhension des sections 5.1, 5.2 et 6.1.13 de ce mémoire.

Pour récupérer des données publiées selon les standards du Web sémantique, on privilégie en général trois méthodes différentes : le téléchargement de *dumps,* le fait de parcourir les URI et, finalement, la formulation de requêtes sur un point d'accès *SPARQL Protocol and RDF Query Language* (SPARQL) (*endpoint*) (Bermès, Isaac et Poupeau, 2013). D'abord, les *dumps* permettent le téléchargement massif des données en RDF. En général, il s'agit simplement d'un lien cliquable qui déclenche un téléchargement automatique. C'est l'une des manières les moins complexes d'avoir accès aux données, car celles-ci sont simplement « déversées » par l'institution.

Ensuite, le fait de parcourir les URI consiste à naviguer à partir d'un URI vers d'autres données et ressources représentées aussi par des URI. Ainsi, grâce à la propriété seealso ou à des liens vers des ressources externes, il est possible de suivre ces liens vers d'autres données ou ressources (internes ou externes) qui sont en lien avec l'URI consulté initialement. Par exemple, un usager consulte une application du Web de données proposant des informations au sujet d'auteurs. En accédant à une page sur un auteur donné, il est possible que certaines informations soient manquantes ou que l'usager souhaite obtenir de l'information au sujet d'auteurs du même courant littéraire. Des liens pourraient donc être proposés permettant à l'usager d'accéder à des ressources externes sur l'auteur ou encore à une liste d'auteurs semblables au sein même de l'application.

Puis, un point d'accès SPARQL est un service qui respecte le protocole SPARQL et qui présente la possibilité d'effectuer des requêtes dans une base de données de triplets. Ces requêtes peuvent être effectuées par les machines ou par les humains, via une page HTML. Ainsi, le

langage de requête SPARQL permet d'effectuer des recherches au sein de données encodées en RDF, de les récupérer et de les manipuler, et ce en temps réel. SPARQL est relativement semblable à *Structured Query Language* (SQL) qui, pour sa part, permet d'effectuer des requêtes dans des bases de données relationnelles (Bermès, Isaac et Poupeau, 2013; Stuart, 2011). En proposant un point d'accès SPARQL, on permet donc aux usagers d'obtenir uniquement les données voulues et ainsi éviter de devoir parcourir l'entièreté des données contenues dans un *dump* et sans avoir besoin de connaître les URI. Bien que l'apprentissage du langage de requête SPARQL ne soit pas excessivement difficile, il peut paraître un peu plus complexe pour certains usagers étant donné la formulation requise pour effectuer des recherches. Une requête SPARQL comprend la déclaration de préfixes, la définition du jeu de données (facultatif), la définition des résultats souhaités, la définition des conditions relatives à la requête et, finalement, les modificateurs de résultats qui permettent, par exemple, d'obtenir les résultats en ordre alphabétique (facultatif) (Feigenbaum et Prud'hommeaux, 2013). Les résultats peuvent être retournés dans plusieurs formats différents tels que XML, JSON, CSV, RDF ou HTML. Il existe quatre modes d'interrogation de SPARQL (Bermès, Isaac et Poupeau, 2013) :

1. SELECT permet d'obtenir des résultats suite à une recherche;
2. CONSTRUCT permet de construire un graphe ou de dégager un sous-ensemble du graphe;
3. ASK permet de vérifier s'il existe au moins une réponse à une requête donnée;
4. DESCRIBE permet, à partir du URI d'une ressource, d'obtenir le graphe la décrivant.

À l'aide de la figure 7, nous présenterons uniquement un exemple d'une requête de type SELECT, car il s'agit de la plus utilisée. Nous allons interroger le point d'accès SPARQL de DBpedia[19] afin d'obtenir la liste des pays dont la population est inférieure à 100 000 habitants[20].

[19] http://dbpedia.org/sparql

[20] Exemple adapté de Feigenbaum et Prud'hommeaux (2013)

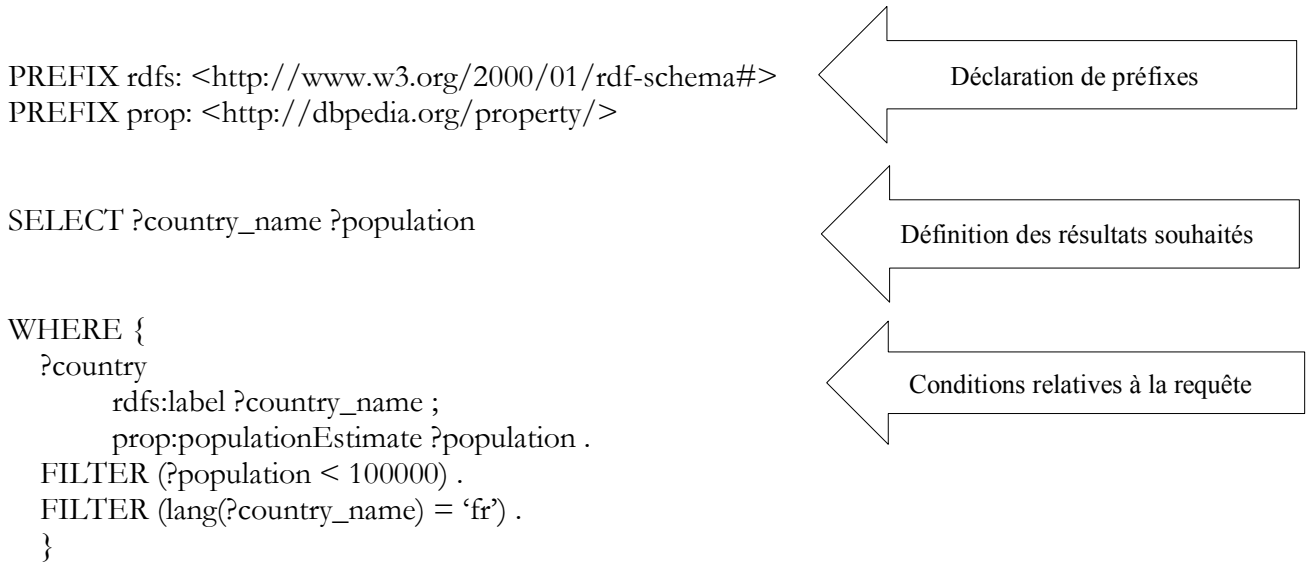


Figure 7. Exemple d'une requête SPARQL de type SELECT

Comme nous l'avons vu à la section 2.5.4, on utilise aussi des espaces de noms lors d'une requête SPARQL. Ici, on constate dans la description une référence aux espaces de noms "http://www.w3.org/2000/01/rdf-schema#" et "http://dbpedia.org/property/". On obtient, suite à cette requête dans un entrepôt de données particulier, un tableau (voir figure 8) présentant la liste des pays dont la population est inférieure à 100 000 habitants. Ces données correspondent à celles qui se trouvent dans l'entrepôt de données en question, soit DBpedia[21].

[21] Notons que le pays Akrotiri et Dhekelia revient deux fois dans le tableau. Une première fois où l'on indique que la population est de 7 700 habitants et une seconde fois où l'on indique que la population est de 8 000 habitants. La distinction vient du fait que la population du pays est composée de 7 700 Chypriotes et de 8 000 militaires britanniques et leurs familles.

country_name	population
"Îles Cocos"@fr	596
"Îles Komandorski"@fr	613
"Terres australes et antarctiques françaises"@fr	140
"Mohéli"@fr	38000
"Akrotiri et Dhekelia"@fr	7700
"Akrotiri et Dhekelia"@fr	8000
"Liberland"@fr	0
"Saint-Barthélemy (Antilles françaises)"@fr	35906
"Samoa américaines"@fr	57345
"Territoire britannique de l'océan Indien"@fr	3000
"Îles Malouines"@fr	2932
"Kerman"@fr	2011
"Saint-Martin (Royaume des Pays-Bas)"@fr	37429
"Géorgie du Sud-et-les Îles Sandwich du Sud"@fr	30
"Wallis (île)"@fr	10731
"Åland"@fr	28666

Figure 8. Capture d'écran d'une partie des résultats obtenus suite à la requête SPARQL présentée à la figure 7

2.6 Vocabulaires et ontologies

Cette section présente les notions de vocabulaires et ontologies utilisées pour les descriptions dans le Web sémantique, nécessaires à la compréhension des sections 3.2.2, 5.1, 6.1.5, 6.1.8 et 6.1.10 de ce mémoire.

Dans le Web sémantique, les ontologies définissent les concepts et les relations afin de classifier l'utilisation de certains termes dans une application donnée, caractériser les relations entre les ressources et mettre sur pied différentes contraintes quant à l'utilisation d'un terme donné (W3C, 2015). Selon Noy et McGuinness (2001, p. 1), « An ontology defines a common vocabulary for researchers who need to share information in a domain. It includes machine-interpretable definitions of basic concepts in the domain and relations among them. » Elles permettent donc de décrire les « choses » de la façon la moins ambigüe possible et doivent être

comprises à la fois par les humains et les machines pour, par la suite, permettre par exemple de filtrer ou classifier l'information (Vatant, 2008). Selon Tom Gruber (s.d.), les ontologies, dans le contexte des sciences de l'information, sont définies comme des règles permettant de modéliser un domaine ou un champ de connaissance. Les ontologies donnent donc du sens aux données structurées, ce qui fait en sorte qu'elles sont utilisées afin d'intégrer des bases de données hétérogènes, de permettre l'interopérabilité entre systèmes et d'apporter des spécifications en ce qui a trait aux interfaces de services axés sur la connaissance. Étant donné que l'objectif est que l'information puisse être « comprise » et traitée de manière utile par les machines, il est nécessaire d'utiliser des vocabulaires qui leur permettront d'interpréter adéquatement les triplets et les relations entre ceux-ci.

Les ontologies doivent être exprimées en RDF, en utilisant les termes (classes et propriétés) proposés par RDFS et OWL. Tout utilisateur du Web peut publier sa propre ontologie, mais il est toujours recommandé de se baser sur des ontologies et vocabulaires plus reconnus afin d'assurer une meilleure interopérabilité. Notons qu'il existe des centaines d'ontologies et de vocabulaires pour différents domaines d'étude et de recherche. Nous en présentons quelques-uns ci-dessous : SKOS (pour l'organisation des connaissances), FOAF (pour les personnes et les liens entre elles), ainsi que le Dublin Core.

2.6.1 Simple Knowledge Organization System (SKOS)

Simple Knowledge Organization System (SKOS) est un langage de définition et de déclaration de vocabulaires contrôlés qui se prête particulièrement bien à la modélisation de taxonomies, de thésaurus, de schémas de classification, de listes de vedettes-matière, de terminologies et de plusieurs autres types de vocabulaires contrôlés (DeWeese et Segal, 2015; Mikhalenko, 2005). Ce vocabulaire permet la représentation de la structure et du contenu de schémas conceptuels qui caractérisent le plus souvent certains types de données. Ainsi, on retrouve des informations terminologiques (skos:prefLabel, skos:altLabel), des liens sémantiques entre les concepts (skos:broader, skos:narrower, skos:related) ainsi que des notes et commentaires (skos:scopeNote, skos:definition) (Isaac, 2012; Vatant, 2008). SKOS est entre autres intéressant

en ce qui a trait aux problématiques que peuvent occasionner les différentes langues utilisées sur le Web ainsi qu'aux différents termes qui peuvent être utilisés pour référer à un même concept. Notons que SKOS est souvent utilisé parallèlement à OWL, ce qui peut parfois créer des confusions. Par exemple, la propriété skos:exactMatch peut être utilisée à la place de la propriété owl:sameAs mentionnée précédemment (section 2.5.3) et skos:Concept est une instance de owl:Class. Il est préférable d'illustrer l'application de cette ontologie par un exemple simple. Si la ressource que l'on veut décrire est le concept de la boisson qu'est le café, on pourrait le décrire ainsi, en Turtle :

```
@prefix skos: <http://www.w3.org/2004/02/skos/core#> .

<http://www.exemple.com/concept/café>
        a skos:Concept ;
        skos:definition "Infusion préparée à partir de fèves de caféier torréfiées et moulues"
        ;
        skos:prefLabel "café" ;
        skos:altLabel "caoua" ;
        skos:broader <http://www.exemple.com/concept/boisson_chaude> ;
        skos:narrower <http://www.exemple.com/concept/cafe_au_lait> .
```

On remarque donc qu'on fait référence à un concept en particulier qui a sa définition propre, mais que certains appellent aussi « caoua » et qui fait partie de la plus grande classe qu'est « boisson chaude » tandis que « café au lait » fait partie pour sa part du concept « café ». SKOS joue donc un rôle important dans l'encodage des descriptions de manière à ce que les machines puissent comprendre la sémantique derrière les ressources, ce qui s'avère particulièrement intéressant pour le domaine des bibliothèques.

2.6.2 Friend of a friend (FOAF)

Ensuite, FOAF est un vocabulaire qui permet la construction d'ontologies dont l'objectif est de décrire des personnes sur le Web ainsi que les relations que celles-ci entretiennent avec d'autres personnes ou des documents. Conçue par Dan Brickley et Libby Miller au début des

années 2000, elle était principalement associée au Web social, car son utilisation était surtout répandue dans ce contexte. En effet, FOAF permettait aux individus de créer des pages à leur sujet dont l'information était accessible à tous. Par exemple, on retrouve dans ce langage de description d'ontologies des classes comme foaf:knows, foaf:Person, foaf:Organization, foaf:Document et des propriétés telles que foaf:made, foaf:mbox, etc. (Bermès, Isaac et Poupeau, 2013; Stuart, 2011). FOAF permet donc à chacun de créer un URI pour se représenter soi-même et ainsi permettre la création de liens d'une personne à une autre puis, éventuellement, construire un graphe de son propre réseau social. Voici un exemple d'une page FOAF simple dans la syntaxe XML/RDF :

```
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
            xmlns:foaf="http://xmlns.com/foaf/0.1/">
   <foaf:Person>
         <foaf:name>Jane Doe</foaf:name>
         <foaf:mbox rdf:resource="mailto:janedoe@exemple.com"/>
         <foaf:homepage rdf:resource="www.janedoe.com"/>
  </foaf:Person>
</rdf:RDF>
```

Cet exemple démontre donc qu'une personne, dont le nom (foaf:name) est Jane Doe, a comme adresse courriel (foaf:mbox) le janedoe@exemple.com et l'URL de son site Internet personnel (foaf:homepage) est le www.janedoe.com. À première vue, cette ontologie ne semble pas appropriée afin de décrire des auteurs d'œuvres littéraires par exemple, mais avec son utilisation grandissante dans le cadre du Web sémantique, les développeurs ont créé une distinction claire entre les propriétés de base (*FOAF Core*) et les propriétés utilisées dans le cadre du Web social (*Social Web*). Les propriétés de base décrivent les caractéristiques relatives aux personnes et aux communautés qui ne dépendent pas du contexte technologique ou temporel (Brickley et Miller, 2014). Ainsi, on y retrouve la définition de classes telles que Agent, Organization et Group. Comme la plupart des ontologies, elle est conçue pour être utilisée de pair avec d'autres vocabulaires.

2.6.3 Dublin Core Metadata Initiative (DCMI)

Le DCMI est un schéma de métadonnées normalisé composé de classes et de propriétés qui permettent de décrire des ressources documentaires (Bermès, Isaac et Poupeau, 2013). Il comprend 15 éléments de base (voir tableau 2) qui peuvent être raffinés grâce aux éléments du Dublin Core qualifié et chacune des propriétés et des classes peuvent être identifiées par un URI (DCMI, 2012). Ce schéma est donc compatible avec le Web de données et il est possible de l'utiliser pour décrire des ressources, comme nous l'avons vu précédemment. Le DCMI est souvent utilisé dans le domaine de l'information, car il se démarque par sa simplicité ainsi que sa flexibilité.

Tableau 2. Éléments du Dublin Core non qualifié

Nom de l'élément	**Nom de l'élément en anglais**	**Description**
Contributeur	*Contributor*	Une personne ou une organisation qui a contribué à la création de la ressource
Couverture	*Coverage*	Couverture spatiale ou temporelle
Créateur	*Creator*	Une personne ou une organisation responsable de la création de la ressource
Date	*Date*	Un moment ou une période associés à un événement dans le cycle de vie de la ressource
Description	*Description*	Résumé ou description du contenu de la ressource
Droits	*Rights*	Informations à propos des droits relatifs à la ressource
Éditeur	*Publisher*	Une personne ou une organisation responsable de la publication de la ressource
Format	*Format*	Le format ou les dimensions de la ressource
Identifiant	*Identifier*	Identificateur non ambigu
Langue	*Language*	La langue de la ressource

Relation	*Relation*	Le lien avec une ou plusieurs autres ressources
Source	*Source*	Ressource de laquelle dérive la ressource décrite
Sujet	*Subject*	Sujet(s) dont traite la ressource
Titre	*Title*	Le nom donné à la ressource
Type de ressource	*Type*	La nature ou le genre de la ressource

Comme mentionné précédemment, un nombre très important d'ontologies et de vocabulaires sont présentement disponibles. Malgré cette situation, il semble y avoir tout de même un problème en ce qui a trait à la description d'objets et de documents patrimoniaux (matériels et immatériels). L'*Europeana Data Model* (EDM) est utilisé par plusieurs institutions, car il a répondu dans une certaine mesure à certains de ces besoins en réutilisant plusieurs espaces de noms tels que *OAI Object Reuse and Exchange* (ORE), utilisé pour définir les standards de description et d'agrégation de ressources numériques, SKOS, Dublin Core, FOAF, etc. À cette utilisation, EDM ajoute ses propres propriétés et classes afin de perfectionner la description. Son objectif de base est de permettre une interopérabilité pour les ressources patrimoniales numérisées, donc de décrire des documents provenant de bibliothèques, de musées et d'institutions archivistiques (Bermès, Isaac et Poupeau, 2013). Cependant, il est important de souligner le fait que le choix des ontologies et des vocabulaires utilisés dépend majoritairement des objectifs de l'institution ainsi que des jeux de données que celle-ci souhaite décrire.

2.7 Données structurées internes

Cette section présente les données structurées internes, aussi appelées microdonnées ou microformats, nécessaires à la compréhension des sections 3.2.6, 5.2.2 et 6.1.13 de ce mémoire.

Une autre composante de la stratégie de visibilité des données sur le Web est celle des données structurées internes. Elles visent à ajouter une nouvelle couche sémantique à l'indexation des pages Web, lisibles et interprétables pour les moteurs de recherche. Elles sont

codées directement dans le code HTML et augmentent donc les chances de repérage et la performance. Cette application est un bon exemple d'utilisation de standards qui ne se limitent plus à ceux utilisés par le domaine des bibliothèques et des institutions documentaires, mais qui permettent une intégration au sein de l'architecture du Web dans son ensemble (Bermès, Isaac et Poupeau, 2013). L'une des conséquences concrètes de l'ajout de données structurées internes dans le code HTML de pages Web est l'apparition d'informations supplémentaires à même les résultats de recherche. Par exemple, si un usager effectue une recherche dans le moteur de recherche Google pour obtenir une recette de lasagne, il pourra constater que les informations suivantes seront visibles dans certains résultats de recherche : une note sur 5 qui représente la moyenne des notes données par les usagers, le nombre de votes, le temps nécessaire pour faire la recette et le nombre de calories par portion (voir figure 9).

Figure 9. Capture d'écran d'une partie des résultats de recherche dans Google pour le terme « lasagne »

Un autre exemple est celui de l'ajout d'informations liées aux heures d'ouverture, aux coordonnées, aux événements à venir et plusieurs autres informations pour des musées, salles de spectacles, bibliothèques, etc. (voir figure 10).

Figure 10. Capture d'écran d'une partie des résultats de recherche dans Google pour les termes « centre bell »

Ces deux exemples démontrent bien l'apport de l'ajout des données structurées internes pour les organisations et les usagers.

En juin 2011, les moteurs de recherche Google, Bing et Yahoo! ont annoncé l'initiative schema.org[22], dont l'objectif était la création, la maintenance et la promotion d'un vocabulaire commun pour décrire les données structurées et la nature des ressources dans les pages Web (Bermès, Isaac et Poupeau, 2013; Guha, 2011; Guha, Brickley et Macbeth, 2016). Le modèle est relativement générique et ne prétend pas couvrir tous les concepts, mais s'inspire tout de même de travaux précédents tels que microformats, les microdonnées et FOAF (schema.org, s.d.). Schema.org propose 638 classes organisées hiérarchiquement et 965 relations pouvant

[22] http://schema.org/

avoir plus d'un domaine et plus d'une portée (Guha, Brickley et Macbeth, 2016). Grâce à la participation et la collaboration de la communauté, schema.org a su s'adapter et évoluer au cours des années. Du côté du domaine des bibliothèques, on note la présence du groupe du W3C nommé *Schema Bib Extend Community Group*[23] dont l'objectif est de proposer des extensions pour schema.org qui amélioreraient la représentation d'informations bibliographiques dans le balisage des pages Web.

En général, on utilise schema.org avec les formats *Resource Description Framework in Attributes* (RDFa), Microdata ou *JavaScript Object Notation for Linked Data* (JSON-LD). Un format n'est pas meilleur qu'un autre et chacun présente des avantages et des désavantages. D'abord, JSON-LD est un standard du W3C qui prend de plus en plus d'ampleur, principalement parce que Google en encourage l'utilisation et l'utilise dans le cadre du projet Google Knowledge Graph (Google, s.d.). De plus, JSON-LD est basé sur le format de données JSON, déjà grandement utilisé, ce qui permet à un plus grand nombre de se l'approprier. Ensuite, RDFa, une recommandation du W3C, est intéressant dans le cadre du Web sémantique dans la mesure où sa syntaxe est en RDF. Ce format est un ensemble d'attributs et d'éléments que l'on peut ajouter à la syntaxe HTML déjà existante. De plus, RDFa permet l'utilisation d'espaces de nom, donc de travailler à l'aide de plusieurs vocabulaires à la fois.

Les définitions des concepts qui ont été décrits dans ce chapitre permettent une meilleure compréhension des enjeux qui se présentent aujourd'hui pour les professionnels de l'information. On comprend l'ampleur que peut prendre la formation des acteurs du domaine et le temps que celle-ci peut nécessiter, mais aussi l'ouverture aux changements et l'évolution inévitable auxquelles ceux-ci devront faire face s'ils souhaitent jouer un rôle au sein de cette nouvelle extension du Web. Notons que tous les concepts composant le Web sémantique et le Web de données n'ont pas été présentés, mais ces descriptions seront utiles pour la bonne compréhension des enjeux et des recommandations présentés dans les pages à venir.

[23] https://www.w3.org/community/schemabibex/

3. Pertinence et défis pour les professionnels de l'information

Dans le domaine des sciences de l'information, le Web de données présente plusieurs avantages, à la fois étant donné la familiarité des professionnels de l'information avec les formats standardisés et les données structurées, mais aussi grâce aux changements évolutifs découlant des nouvelles technologies auxquels ceux-ci ont dû et doivent toujours faire face. De plus, les objectifs du Web de données sont intimement liés à ceux des institutions documentaires. Ce chapitre est séparé en deux parties, soit la présentation du lien entre les bibliothèques et les standards et les avantages et défis que présente le Web de données pour ces institutions.

3.1 Bibliothèques et standards

Les professionnels de l'information sont formés de façon à être aptes à analyser de l'information structurée, à effectuer la description de contenu en suivant des standards et des normes et à participer à l'évolution et à la définition de formats informatiques qui permettront de conserver, manipuler et échanger les données (Bermès, Isaac et Poupeau, 2011). Bien que l'objectif ne soit pas de faire des professionnels de l'information des informaticiens, il est important de miser sur cette sensibilité quant à l'encodage des données structurées déjà acquises afin de permettre un changement et une évolution des catalogues de bibliothèques. L'utilisation de standards, de normes et de principes rigoureux au sein de bibliothèques à travers le monde est basée sur le travail de pionniers tels que Charles Cutter, Melville Dewey, Paul Otlet, S.R. Ranganathan et Seymour Lubetzky (Alemu, Stevens, Ross et Chandler, 2012; Denton, 2007). Leur travail a mené la communauté de bibliothécaires à la publication de ces standards et à des ententes communes visant à l'interopérabilité des données comme l'Anglo-American Cataloguing Rules (AACR) en 1967, le standard MAchine-Readable Cataloguing (MARC) à la fin des années 1960s, l'International Standard Bibliographic Description (ISBD) en 1971, le Functional Requirements for Bibliographic Records (FRBR) en 1996 et le Resource Description and Access (RDA) en 2010. Tous ces standards et principes sont en général maîtrisés par la communauté des professionnels de l'information et ont eu un impact sur la vision de ce que devaient et doivent être les catalogues de bibliothèques ainsi que leur objectif. Afin de

comprendre la pertinence du Web sémantique et du Web de données pour le domaine des bibliothèques, il est d'abord nécessaire d'accepter les changements auxquels les professionnels de l'information doivent faire face, non seulement au niveau de la gestion des documents physiques et numériques, mais aussi en ce qui a trait aux comportements informationnels des usagers qui ont été radicalement modifiés suite à l'arrivée du Web. Ces standards nommés préalablement gardent leur pertinence, mais il est primordial de mettre de l'avant l'aspect d'interopérabilité, non seulement dans le contexte bibliothéconomique, mais aussi dans le contexte infiniment plus large qu'est le Web.

Les catalogues informatisés ont donné naissance au format MARC dont l'objectif était de permettre de dupliquer les notices d'un catalogue à un autre et, ainsi, éviter la duplication de l'effort de catalogage d'une institution à une autre (Bermès, Isaac et Poupeau, 2013). Le protocole Z39.50 permet à des systèmes intégrés de gestion de bibliothèques (SIGB) de communiquer entre elles. Il offre la possibilité d'interroger plusieurs catalogues de bibliothèques simultanément et est basé sur l'interopérabilité que permet le format MARC. Cependant, les catalogues en ligne restent dans le « Web profond », c'est-à-dire qu'ils sont des silos dont les données sont isolées et ne sont accessibles que dans la mesure où un usager est au courant de leur existence (Alemu, Stevens, Ross et Chandler, 2012; Hannemann et Kett, 2010; Stuart, 2011). Les données qui composent le catalogue d'une bibliothèque ne font donc pas partie de l'espace global d'information qu'est le Web. Les bibliothèques sont de plus en plus composées de documents numériques et cette réalité est en soi une occasion de réévaluer le catalogue et son objectif et de mettre de l'avant les compétences informatiques et la capacité d'adaptation des professionnels de l'information.

3.2 Bibliothèques et Web de données : avantages et défis

Les avantages et défis du Web de données pour les professionnels de l'information présentés ci-dessous sont basés sur le rapport final du groupe d'incubation « Bibliothèques et Web de données » de Baker et al. (2011). La mission de ce groupe s'est déroulée entre mai 2010 et août 2011 et son objectif était d'identifier des pistes de collaboration pour l'avenir et de

contribuer à l'amélioration de l'interopérabilité entre les données de bibliothèques sur le Web. Le rapport final donne des pistes sur l'utilisation des standards du Web sémantique ainsi que sur les fondements du Web de données afin que les données normalisées et stockées par les bibliothèques aient un plus grand rayonnement ainsi qu'une possibilité de réutilisation, et ce, à l'échelle du Web (Baker et al., 2011). D'un côté, les avantages présentés par le groupe sont les suivants : la réutilisation des données, le multilinguisme, l'enrichissement des données, l'interopérabilité, la sérendipité, la présence sur le Web et la pérennité. Nous les abordons plus en détail ci-dessous. D'un autre, les défis soulevés par le groupe sont les suivants :

- Les normes bibliographiques sont conçues seulement pour la communauté des bibliothèques ;
    - Les formats utilisés par les professionnels de l'information sont spécifiques à leur domaine, ce qui rend la tâche difficile lorsqu'on tente de s'associer à d'autres données qui ont été créées et qui sont maintenues par des communautés provenant d'autres domaines.
- La communauté des bibliothèques et celle du Web sémantique utilisent des termes différents pour exprimer les mêmes concepts ;
    - Il est difficile d'atteindre une certaine uniformisation et une compréhension commune des mêmes concepts lorsque les communautés ont chacune leur propre vocabulaire pour les décrire.
- Les évolutions technologiques en bibliothèque dépendent de systèmes commerciaux ;
    - Le nombre de fournisseurs pour les bibliothèques est limité et ceux-ci seulement permettent l'accès à des solutions technologiques qui répondent aux besoins précis du domaine absolument nécessaire au bon fonctionnement des institutions. Les bibliothèques sont donc dépendantes des changements technologiques apportés par ces fournisseurs.
- Les droits de propriété peuvent être complexes.
    - Les droits peuvent varier d'un territoire à un autre et une certaine incertitude peut s'installer en ce qui a trait aux droits de publication de certaines ressources et de leurs données.

3.2.1 Réutilisation des données

Les données liées, et surtout les données liées ouvertes, peuvent être partagées et réutilisées facilement. C'est l'objectif même du Web de données : permettre la réutilisation de données par les moteurs de recherche et les agents. Comme mentionné précédemment, les bibliothèques ont déjà mis en place des protocoles visant à partager les données et à favoriser l'interopérabilité (p. ex. Z39.50 et *Open Archives Initiative Protocol for Metadata Harvesting* (OAI-PMH)[24] sur lequel nous reviendrons). Cependant, de tels protocoles ont un grand nombre de désavantages qui ont un impact sur l'interaction entre les différents domaines, qui ne permettent pas certaines opérations et qui n'ouvrent pas nécessairement les portes à toute l'étendue des métadonnées des bibliothèques (Bermès, Isaac et Poupeau, 2013; Stuart, 2011). Ainsi, il ne s'agit pas d'un changement radical quant au partage de ces métadonnées, mais d'une extension naturelle du principe de catalogue de bibliothèque (Coyle, 2010).

3.2.2 Multilinguisme

Ensuite, le Web de données permet la mise en place de fonctionnalités multilingues qui permettent la création de nouveaux services aux utilisateurs. Étant donné que les concepts sont représentés par des URI, l'ambigüité par rapport à ceux-ci est largement diminuée. Grâce au vocabulaire SKOS mentionné précédemment, il est aussi possible de lier à un concept donné plusieurs mots provenant de langues différentes, permettant ainsi d'accéder à la ressource recherchée, même si celle-ci est dans une autre langue. Par exemple, si le concept choisi est « Guerre mondiale, 1939-1945 », voici une façon de spécifier les expressions alternatives pour parler du même concept, en anglais et en français, grâce à l'ontologie SKOS, dans la syntaxe de sérialisation Turtle :

[24] https://www.openarchives.org/pmh/

```
@prefix skos: <http://www.w3.org/2004/02/skos/core#> .

<http://exemple.com/concept#guerremondiale3945>
        a skos:Concept ;
        skos:prefLabel "Guerre mondiale, 1939-1945"@fr, "World War, 1939-1945"@en ;
        skos:altLabel "2e guerre mondiale"@fr, "Deuxième guerre mondiale"@fr,
        "Seconde guerre mondiale"@fr, "Second World War"@en, "World War 2"@en,
        "WWII"@en, "World War Two"@en .
```

Comme on peut le constater, on utilise skos:prefLabel pour indiquer l'expression préconisée en français et en anglais (Guerre mondiale, 1939-1945 et World War, 1939-1945) et skos:altLabel pour indiquer les expressions alternatives dans les deux langues (2e guerre mondiale, Deuxième guerre mondiale, Seconde guerre mondiale, Second World War, World War 2, WWII et World War Two).

3.2.3 Enrichissement des données

Le Web de données permet donc aussi un enrichissement des données de la bibliothèque grâce à un apport externe. En effet, étant donné que les œuvres, lieux, personnes, événements, sujets ou concepts sont représentés par des URI pérennes, le référencement des ressources d'une bibliothèque peut être enrichi grâce à d'autres jeux de données et, ainsi, améliorer la qualité de ses métadonnées descriptives. Si l'on retrouve sur un serveur des informations concernant la Guerre mondiale de 1939-1945 et que sur un autre on en retrouve d'autres complémentaires, celles-ci seront reliées entre elles et permettront de créer un seul graphe dont le point commun sera ce concept. Il sera donc possible de naviguer d'une ressource à l'autre et, ainsi, d'accéder à toutes les informations pertinentes et recherchées. Par la suite, une bibliothèque pourra construire des liens de sa ressource vers les autres ressources traitant du même concept, s'assurant donc que les métadonnées descriptives de la ressource soient complètes.

Par exemple, une bibliothèque qui aurait un document donné pourrait effectuer un lien vers la traduction de ce document qui pourrait se trouver dans une autre institution. Ainsi, la plupart des contributions sont intéressantes, principalement lorsque les fournisseurs de données ont l'occasion de proposer des parties de leurs données sous la forme de déclarations.

Contrairement à la pratique établie où les données sont partagées sous la forme de notices, il s'agit ici de la possibilité de créer des graphes sur des ressources données dont les informations proviennent de sources différentes. Par le fait même, on constate la possibilité de diminuer la redondance en ce qui a trait aux descriptions bibliographiques et aux métadonnées. Il s'agit donc d'une occasion d'optimiser les coûts reliés aux activités de description des fonds. Conséquemment, les bibliothèques peuvent partager entre elles les données qui sont déjà accessibles via des sources fiables et ainsi éviter la duplication du travail (Alemu, Stevens, Ross et Chandler, 2012; Tillett, 2013). Les professionnels responsables du catalogage pourront donc mettre de l'avant leur expertise dans leur domaine, plutôt que de retravailler des descriptions et des notices déjà faites par d'autres (Baker et al., 2012).

3.2.4 Interopérabilité

Le fait d'avoir un format universel pour le partage de données est aussi intéressant dans la mesure où cela permettrait une meilleure interopérabilité d'une institution à une autre, mais aussi d'une institution à d'autres instances publiant leurs données ouvertes et ce, peu importe le système utilisé. Les technologies développées pour la gestion des données de bibliothèques sont présentement produites par les fabricants de SIGB et les normes bibliographiques sont donc conçues uniquement pour la communauté des professionnels de la bibliothèque, ce qui limite les possibilités de partage qui s'offrent aux bibliothèques et contribue largement à l'isolement de leurs données. Aussi, les petites institutions dont les dépenses sont plus limitées obtiendraient du même coup une plus grande visibilité grâce à une interaction plus importante avec d'autres instances (Baker et al., 2011).

3.2.5 Sérendipité

Ensuite, comme c'est le cas pour le Web en permettant de naviguer d'un document à l'autre ou d'une page Web à l'autre grâce aux liens hypertextes, le Web de données permettra aux usagers de naviguer en suivant le même modèle, mais de données en données. On constate donc que le modèle favorise la sérendipité, c'est-à-dire le fait de faire une découverte de façon

inattendue. Le terme a été utilisé pour la première fois par l'historien anglais Horace Walpole en 1754 (Merton et Barber, 2004, cité dans Alemu, Stevens, Ross et Chandler, 2012). Le phénomène de sérendipité attire de plus en plus l'attention de chercheurs du domaine des sciences de l'information, car la plupart des modèles de comportements informationnels n'en tiennent pas compte jusqu'à maintenant (Alemu, Stevens, Ross et Chandler, 2012). Foster et Ford (2003) ont démontré pour leur part que la sérendipité était un sujet difficile à étudier étant donné qu'il s'agit d'un concept qui, par définition, ne peut être affecté par un contrôle ou une prédiction. Cependant, avec la mutation des services des bibliothèques vers le numérique, les usagers perdraient des occasions de faire des découvertes inattendues étant donné qu'ils n'ont plus accès de la même manière aux étagères de bibliothèques qui leur permettaient auparavant de consulter les livres se trouvant à proximité du document recherché et qui traiteraient du même sujet. Notons cependant l'existence d'« étagères virtuelles » qui permettent de réaliser ce type de recherche d'information dans un contexte numérique. Par exemple, les Bibliothèques de l'Université de Montréal proposent un tel outil au sein de leur catalogue en ligne *Atrium* (voir figure 11).

Figure 11. Étagère virtuelle disponible dans le catalogue en ligne *Atrium* des Bibliothèques de l'Université de Montréal

Ainsi, mis à part cet exemple peu commun, la question est à savoir si les bibliothèques numériques peuvent mener au même phénomène de sérendipité que celui vécu lors d'une

recherche dans la bibliothèque physique. Étant donné que l'accès aux ressources dans les bibliothèques numériques dépend des métadonnées qui les décrivent, le Web de données permettrait de naviguer d'une ressource à l'autre, même si ces ressources ne font pas partie du même catalogue de bibliothèque. Contrairement aux catalogues en ligne qui sont basés sur le fait d'effectuer une recherche pour obtenir des résultats (par exemple le nom d'un auteur ou le titre d'un document), le Web des données ouvre la porte à des recherches plus larges. En effet, grâce à ces technologies, l'usager peut faire des découvertes inattendues et naviguer d'une bibliothèque numérique à une autre ou encore d'une bibliothèque numérique à des sources externes telles que Wikipedia ou GeoNames (une base de données géographiques) par exemple (Alemu, Stevens, Ross et Chandler, 2012; Baker et al., 2011; Coyle, 2010). Nous présenterons un exemple à la section 5.1.1.1. On constate du même coup la possibilité d'effectuer des recherches fédérées ainsi que des recherches interdisciplinaires plus performantes, qui iraient chercher l'information dans un plus grand nombre de bases de données ouvertes.

3.2.6 Présence sur le Web

L'utilisation grandissante de moteurs de recherche tels que Google pour trouver de l'information sur un sujet donné a un impact majeur sur la perception qu'ont les usagers du rôle du bibliothécaire. D'ailleurs, Google ayant fait l'acquisition de Freebase (une base de données collaborative) en 2010 (Google, 2010), il s'agit d'une démonstration de la compréhension du Web de données par ce moteur de recherche. Google a par la suite migré les données de Freebase vers sa nouvelle plateforme, Wikidata[25]. Nous y reviendrons à la section 5.1.1.5. Les nouvelles technologies relatives au Web ont été accueillies de façon positive par la communauté des professionnels de l'information pour plusieurs raisons telles que l'arrivée de nouveaux services comme l'accès rapide à des articles scientifiques, de nouveaux modes de communications avec les usagers (p. ex. clavardage), l'accès à des catalogues de bibliothèques différentes, etc. Cependant, Google, entre autres, a rapidement été considéré comme une menace pour certains professionnels, étant donné que la plupart des usagers jugent que l'information qu'ils y trouvent

[25] https://www.wikidata.org/

est suffisante et qu'ils n'iront pas chercher l'aide pour l'obtenir (Stuart, 2011). Les compétences informationnelles sur le Web des professionnels de l'information sont nettement plus développées, entre autres grâce à leur meilleure compréhension du potentiel que peuvent apporter les caractéristiques d'une recherche avancée, à la capacité d'évaluer la pertinence d'une source et à l'utilisation de la logique booléenne dans une recherche. Ainsi, étant donné que les données de bibliothèques sont présentement isolées du reste du Web, une requête sur un moteur de recherche ne permettra pas d'avoir accès à celles-ci (Coyle, 2010; Gonzalez, 2014; Zeng et Qin, 2008, cité dans Schilling, 2012). Un autre aspect intéressant du Web des données est donc le fait que RDFa et les microdonnées (*microdata*) permettent une présence des données de bibliothèques dans les résultats des moteurs de recherche et, par conséquent, une plus grande visibilité. Les ressources des catalogues de bibliothèques seraient ainsi disponibles pour les usagers du Web, ce qui ferait en sorte que l'expertise des bibliothécaires et professionnels de l'information pourrait à nouveau être mise de l'avant grâce à la qualité et la fiabilité de ces ressources.

3.2.7 Pérennité

Nous avons abordé préalablement la question des URI pérennes (section 2.5.1) et la question de la durée dans le temps de ces technologies est pertinente. La plupart du temps, la vie des technologies est de très courte durée et l'on voit l'apparition de nouveaux formats qui prennent la relève. Dans leur rapport, Baker et al. (2011) soulignent que le fait que les données liées sont caractérisées par le fait qu'elles décrivent la sémantique, et ce, indépendamment des formats et de la syntaxe. Ainsi, même si un nouveau format voit le jour, les données gardent leur signification et sont conséquemment plus pérennes. En effet, les technologies du Web sémantique sont basées sur des formats dont la pérennité est la plus probable tels que .txt et .xml.

En général, la plupart des données retrouvées dans les catalogues de bibliothèques sont présentement encodées en format texte en langage naturel (Baker et al., 2011). Cette situation est problématique dans la mesure où un grand nombre de données qui se trouvent dans les

notices MARC pourraient être en fait représentées par des identifiants qui pourraient par la suite être liés à d'autres ressources (ISBN, noms de personnes et de matières, etc.). Le problème réside donc dans le fait que lorsque les notices sont construites, elles le sont au point de vue local et ne sont pas liées à d'autres instances. Procéder à une modification de l'affichage des données dans les notices peut s'avérer être un travail de longue haleine et relativement coûteux, car il faudrait alors retravailler toutes les notices bibliographiques. Cet aspect a donc un impact sur la décision de s'engager dans le Web de données, car les ressources nécessaires pour procéder à la transition sont importantes.

4. Descriptions générales des bibliothèques nationales

Afin de pouvoir évaluer dans quelle mesure l'implantation d'un projet de Web de données à BAnQ serait réalisable, quelles ressources seraient impliquées pour y parvenir et quelles étapes sont nécessaires à cette mise en œuvre, il est d'abord pertinent de comparer les deux bibliothèques. Celles-ci étant toutes les deux des bibliothèques nationales évoluant au sein de la francophonie, il est intéressant de se pencher sur leur historique et mission, leur organisation, les ressources qui ont été mises en ligne ainsi que l'aspect de diffusion et collaboration, afin d'en dresser des portraits généraux. Il est donc nécessaire de porter une attention particulière à ces aspects afin de bien comprendre les environnements dans lesquels ont été ou pourraient être déployés des projets de Web de données. De plus, la comparaison est pertinente, car elle permet de justifier le choix de la BnF comme source d'analyse et d'éventuellement être apte à identifier quels aspects constituent les forces et les faiblesses de BAnQ en comparaisons avec celles de la BnF.

4.1 Bibliothèque et Archives nationales du Québec (BAnQ)

4.1.1 Historique et mission

Bibliothèque et Archives nationales du Québec est une société d'État québécoise qui a vu le jour d'abord suite à la fusion de la Bibliothèque nationale du Québec (BNQ) et de la Grande Bibliothèque, puis d'une nouvelle fusion avec Archives nationales du Québec (ANQ), en 2006. Créées en 1920, les Archives nationales du Québec avaient alors comme objectif d'assurer principalement la gestion et la conservation des archives du Régime français et de rassembler la documentation traitant de l'histoire du Québec (BAnQ, s.d.-a). En 1961, on rattache les Archives nationales du Québec au nouveau ministère des Affaires culturelles et, en 1967, l'Assemblée nationale du Québec adopte une loi menant à la création de la Bibliothèque nationale du Québec, qui relèvera du même ministère. Puis, 1968 voit naître le Règlement sur le dépôt légal qui impose aux éditeurs de fournir deux exemplaires de chaque œuvre imprimée (livres, brochures, journaux, revues, livres d'artistes et partitions musicales sont alors touchés par ce règlement) (BAnQ, s.d.-a). Suite à la mise en place du dépôt légal, apparait l'année

suivante la Bibliographie du Québec qui répertorie tous les documents publiés au Québec ainsi que des documents publiés à l'étranger qui sont en lien avec le Québec, soit par le sujet, soit par l'auteur. En 1992, l'Assemblée nationale autorise que le dépôt légal soit appliqué à des documents d'autres formats tels que les estampes originales, les affiches, les enregistrements sonores, les logiciels, les documents électroniques, etc. En 1994, BNQ ouvre l'accès gratuit à toutes ses collections grâce à la mise en ligne du catalogue Iris et ANQ fait de même en offrant l'accès aux chercheurs à toutes leurs bases de données sur les archives grâce au système informatique Pistard (Programme informatisé servant au traitement des archives et à la recherche documentaire). Aujourd'hui encore, ces catalogues en ligne portent les mêmes noms et occupent une place importante en ce qui a trait à la diffusion des collections et des fonds d'archives.

En 1996, suite à un manque d'espace, la Ville de Montréal et le gouvernement du Québec se penchent sur la question de la pertinence de relocaliser les collections de BNQ et de la Bibliothèque centrale de Montréal et créent un comité présidé par Clément Richard dont l'objectif est d'évaluer la pertinence de construire une grande bibliothèque publique au Québec (BAnQ, s.d.-a). Un an plus tard, le rapport Richard est déposé et la conclusion indique que la construction de cette bibliothèque est nécessaire. En 2001, le gouvernement du Québec adopte une loi visant à fusionner BNQ et la Grande Bibliothèque du Québec et cette société d'État est nommée Bibliothèque nationale du Québec (BNQ). De 2001 à 2004 se déroule la construction de la Grande Bibliothèque en plein centre du Quartier latin à Montréal. Puis, les modifications apportées à la *Loi sur la Bibliothèque nationale du Québec* et à la *Loi sur les archives* en 2006 donnent naissance à une nouvelle institution, Bibliothèques et Archives nationales du Québec (BAnQ). La mission de BAnQ est triple (BAnQ, s.d.-b).:

1. La conservation et la diffusion du patrimoine ;
2. La diffusion et la promotion du savoir ;
3. La mission dans le domaine des archives.

Le tableau 3, construit suite à la consultation de la page « Mission » du site Web officiel de BAnQ, présente ces missions de façon plus détaillée.

Tableau 3. Missions de BAnQ présentées sur le site officiel de l'institution

Mission	Détails
Conservation et diffusion du patrimoine	- Rassembler, conserver de manière pertinente et diffuser : - le patrimoine documentaire québécois publié et tout document qui s'y rattache et qui présente un intérêt culturel; - tout document relatif au Québec et publié à l'extérieur du Québec.
Diffusion et promotion du savoir	- Offrir un accès démocratique au patrimoine documentaire constitué par ses collections, à la culture et au savoir. - Agir comme catalyseur auprès des institutions documentaires québécoises, contribuant ainsi à l'épanouissement des citoyens. - Poursuite des objectifs suivants : - valoriser la lecture, la recherche et l'enrichissement des connaissances; - promouvoir l'édition québécoise; - faciliter l'autoformation continue; - favoriser l'intégration des nouveaux arrivants; - renforcer la coopération et les échanges entre les bibliothèques; - stimuler la participation québécoise au développement de la bibliothèque virtuelle.
Domaine des archives	- Offrir des services de soutien à la recherche et contribuer au développement et au rayonnement international de l'expertise et du patrimoine documentaire québécois. - Encadrer, soutenir et conseiller les organismes publics en matière de gestion de leurs documents. - Assurer la conservation d'archives publiques, en faciliter l'accès et en favoriser la diffusion. - Promouvoir la conservation et l'accessibilité des archives privées.

4.1.2 Organisation

À la tête de BAnQ se trouve un conseil d'administration présidé par la présidente-directrice générale, Madame Christiane Barbe (de juillet 2014 à aujourd'hui). Celui-ci est composé de 17 membres, dont deux représentants des usagers, élus par les usagers mêmes, qui se rencontrent quatre fois par année afin d'adopter, entre autres, les politiques, les règlementations, le budget et les orientations de BAnQ (BAnQ, s.d.-c). Quatre comités exécutifs appuient le conseil d'administration en proposant des recommandations, soit le comité d'audit, le comité sur les collections et les services de BAnQ, le comité sur les services adaptés et le comité sur les technologies de l'information. Selon le rapport annuel de gestion 2014-2015, le nombre d'employés en poste au 31 mars 2015 s'élevait à 717 pour l'année 2015, soit 43 employés pour le personnel d'encadrement, 219 pour les professionnels et conseillers en ressources humaines et 455 pour les employés de soutien (BAnQ, 2015).

4.1.3 Ressources en ligne

Plusieurs outils de recherche électronique sont présents sur le portail de BAnQ et les principaux seront ici présentés. D'abord, au niveau des bases de données, on retrouve la collection numérique qui permet d'avoir accès librement aux documents numérisés. La collection numérique de BAnQ est composée des documents suivants : annuaires municipaux et commerciaux, archives civiles et judiciaires, archives d'écrivains (Jacques Ferron et Rina Lasnier), fichiers bibliographiques, livres et partitions musicales, ouvrages de référence, publications de communautés autochtones (Publications de la Société Makivik et Publication de l'Institut culturel Avataq : Tumivut), publications gouvernementales, revues et journaux québécois, images, cartes et plans, archives radiophoniques, baladodiffusions, contes pour enfants, musique et enregistrements, vidéos (p. ex., collection Arthur Lamothe, Festival International de Jazz de Montréal, Midis littéraires de la Grande Bibliothèque) (BAnQ, s.d.-d).

Le programme de numérisation présentement en cours a comme projet de diffuser « […] l'ensemble du patrimoine documentaire publié ou archivistique produit au Québec depuis le XVII^{e} siècle ou d'origine étrangère et relatif au Québec » (BAnQ, s.d.-d). La collection numérique se bonifie chaque année et en 2014-2015, 2 549 livres imprimés numérisés, 16 publications en série et 3 307 documents audiovisuels y ont été ajoutés pour un total de 12 982 546 documents des collections patrimoniales en ligne (BAnQ, 2015). Notons que ce total comprend aussi les documents patrimoniaux nés numériques disponibles en ligne. Le tableau 4 présente le nombre de documents patrimoniaux numérisés contre le nombre de documents patrimoniaux nés numériques en date du 31 mars 2015.

Tableau 4. Répartition des documents patrimoniaux numérisés et nés numériques au 31 mars 2015 (BAnQ, 2015)

Nombre de documents patrimoniaux **numérisés** au 31 mars 2015	12 849 451
Nombre de documents patrimoniaux **nés numériques** au 31 mars 2015	133 095
Total – Nombre de documents des collections patrimoniales en ligne	12 982 546

Puis, la répartition des documents patrimoniaux numérisés en ligne par catégories documentaires en date du 31 mars 2015 est illustrée à la figure 12. Il est intéressant de constater quels types de documents sont priorisés en ce qui a trait au processus de numérisation. Les archives textuelles ainsi que les publications en série prennent la plus grande place, alors que les documents audiovisuels et les documents cartographiques sont moins présents.

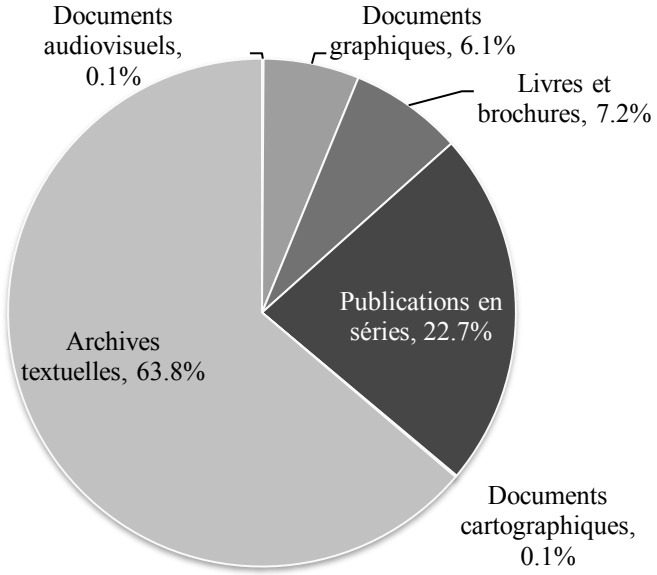


Figure 12. Répartition des documents patrimoniaux numérisés en ligne par catégories documentaires en date du 31 mars 2015, selon le rapport annuel de gestion 2014-2015 (BAnQ, 2015, p.23)

Pistard Archives permet l'accès aux collections d'archives conservées par BAnQ et offre la possibilité d'effectuer des recherches simples ou avancées dans les documents d'archives conservés dans les dix centres d'archives de BAnQ et dans trois services d'archives privés agréés (BAnQ, s.d.-e). On retrouve dans cette base de données des documents numérisés ou non et il est possible de rechercher uniquement les documents qui le sont. Les archives gouvernementales, civiles et notariales, judiciaires et privées s'y retrouvent ainsi que des archives photographiques et iconographiques, cartographiques, architecturales et audiovisuelles.

Ensuite, parmi les outils de recherche, on retrouve le Catalogue des bibliothèques du Québec (CBQ) qui permet d'effectuer des recherches dans plusieurs catalogues de bibliothèques du Québec, de faire une demande de prêt entre bibliothèques en ligne (PEB) et de repérer la bibliothèque la plus près de chez soi, puis être redirigé vers son catalogue (BAnQ, s.d.-g). Cet outil mis en place par BAnQ favorise donc grandement la coopération entre les différentes bibliothèques publiques du Québec et met de l'avant la complémentarité des collections.

Finalement, le catalogue Iris, le catalogue de bibliothèque accessible en ligne, regroupe tous les documents de la Collection universelle et des Collections patrimoniales. En ce qui a trait au nombre d'accès aux catalogues en ligne, on dénombre à 3 033 292 le nombre de visites de la Collection universelle et des collections patrimoniales et à 365 999 visites du Catalogue des fonds et collections d'archives (BAnQ, 2015).

4.1.4 Diffusion et collaboration

Étant donné que le Web sémantique et le Web de données reposent sur des concepts d'ouverture, de coopération et de partage, il est intéressant de se pencher sur les pratiques de diffusion et de collaboration de BAnQ. En raison de sa présence sur les médias sociaux, mais aussi de sa participation au Réseau francophone numérique et aussi au sein de la Bibliothèque numérique mondiale et de la mise sur pied du Service québécois de traitement documentaire, on constate une volonté certaine de diffusion.

4.1.4.1 Médias sociaux

BAnQ assure une présence sur les médias sociaux, soit sur Facebook, Twitter, YouTube et Historypin. Avec 18 454 « J'aime » sur sa page Facebook, 9 575 abonnés à son compte Twitter et 602 abonnés à sa chaîne YouTube en date du 17 mars 2016, BAnQ peut rejoindre un grand nombre d'usagers en peu de temps pour les informer de ses activités et services.

4.1.4.2 Collaboration nationale

Responsable du Service québécois de traitement documentaire (SQTD), BAnQ permet aux bibliothèques québécoises d'avoir accès, grâce à ce service et en s'identifiant, à un grand nombre d'informations pour bonifier leurs catalogues soit les notices catalographiques des documents acquis. On retrouve aussi sur le portail des liens vers des ressources relatives au traitement documentaire comme RDA Toolkit, des plans de classement, WebDewey, un éditeur

de notices, etc. De plus, mentionnons le Catalogue des bibliothèques du Québec (CBQ) qui constitue une implication majeure de BAnQ du point de vue national, présenté à la section 4.1.3.

Ces aspects permettent donc de dresser un portrait général de l'institution et de donner au lecteur une idée des aspects qui la caractérisent. En se basant sur cette description, il est possible de mieux comprendre les différents enjeux auxquels doit faire face BAnQ, mais aussi de constater son implication au sein du domaine. Ces aspects sont repris ci-dessus, mais pour décrire la BnF. On constatera certaines similarités entre les deux institutions, mais aussi le fait que celles-ci n'ont pas nécessairement les mêmes ressources et ne doivent pas faire face aux mêmes difficultés.

4.2 Bibliothèque nationale de France (BnF)

4.2.1 Historique et mission

La Bibliothèque nationale de France, nommée ainsi depuis 1994, est la plus importante bibliothèque de ce pays et se démarque entre autres par le fait qu'elle a hérité des collections royales accumulées depuis Charles V, dit « Charles le Sage » (1338-1380). En effet, après avoir mené le projet de reconstruction du Louvre en 1367, il y aménage une pièce dans une tour pour sa collection de 917 manuscrits. Aussi, il commande des traductions en français de certaines œuvres religieuses, entre autres à Raoul de Presles, mais aussi des traités scientifiques, philosophiques, d'histoire et d'astrologie. Il s'agit de la première *Librairie royale*, confiée au bibliothécaire Gilles de Mallet (Léopold, 1907). Bien que l'objectif ne soit pas alors de rendre cette collection publique, cette initiative de Charles V est le premier pas vers la création d'une bibliothèque nationale. Ensuite, François 1er, avec la multiplication des ouvrages imprimés, décide de mettre sur pied une bibliothèque au château de Blois dans l'objectif de conserver des ouvrages pour les générations futures (Sepetjean et Graff, 2011). L'Ordonnance de Montpellier, signée le 28 décembre 1537, interdit aux librairies et aux imprimeurs de vendre une œuvre imprimée sans en avoir préalablement déposé un exemplaire au château. Cette ordonnance est l'acte de naissance du dépôt légal en France.

Le décret de création n° 94-3 du 3 janvier 1994 stipule que la BnF a comme mission de « […] collecter, conserver, enrichir et communiquer le patrimoine documentaire national » (BnF, 2015-a). Le tableau 5 présente la mission de façon plus détaillée.

Tableau 5. Missions de la BnF présentées sur le site officiel de l'institution (BnF, 2015-a)

Mission	Détails

Collecter, conserver, cataloguer	- Le dépôt légal
	- Enrichir les collections
	- Cataloguer et conserver
	  - Production d'un catalogage de référence
	  - Mettre à disposition des autres bibliothèques des produits bibliographiques pour l'alimentation de leurs propres catalogues

Assurer l'accès du plus grand nombre aux collections	- Numériser pour ouvrir l'accès aux collections
	  - Bibliothèque numérique Gallica
	- Recherche et coopération
	  - Catalogue collectif de France

Ces deux missions sont en partie semblables à celles de BAnQ dans la mesure où elles se basent toutes deux sur la conservation et la diffusion des documents en assurant l'accessibilité à tous et en agissant comme des institutions phares du domaine. Notons cependant que BAnQ est responsable de la conservation des archives nationales du Québec tandis que les archives nationales de France dépendent du ministère de la Culture et de la Communication.

4.2.2 Organisation

Présidée par Bruno Racine (avril 2007 à aujourd'hui), l'organisation de la BnF est basée sur la présence d'un conseil d'administration composé de vingt membres qui se rencontrent trois fois par année pour discuter des orientations de l'institution, du budget, du rapport d'activité, etc. (BnF, 2014-a). Parmi ses membres, le conseil regroupe des représentants des tutelles, des membres du personnel, des personnalités extérieures et deux représentants des lecteurs. Un

conseil scientifique est pour sa part sollicité une fois par an et est composé de dix-sept membres qui se prononcent sur les activités de recherche de l'institution ainsi que sur sa politique scientifique (BnF, 2014-a).

En 2014, 2 393 personnes étaient employées par la BnF (2015-b). La répartition des effectifs diffère grandement du fonctionnement québécois. En effet, on retrouve du personnel d'État, du personnel non titulaire et du personnel non titulaire à temps partiel (260 postes). Ces différents corps d'emploi sont ensuite divisés par catégories (A, B ou C) et le recrutement se fait à l'aide de concours internes ou externes. Les emplois de catégorie A impliquent des tâches de conception et de direction (832 postes), les emplois de catégorie B assurent la gestion et l'application (636 postes) alors que les emplois de catégories C assurent l'exécution (657 postes). On retrouve aussi 8 postes dans des « emplois d'avenir », qui s'apparentent à des stages et qui ont comme objectif de faciliter l'insertion professionnelle et l'accès à l'emploi pour des jeunes en situation difficile (Ministère du Travail, de l'Emploi, de la Formation professionnelle et du Dialogue social, 2013).

4.2.3 Ressources en ligne

Le catalogue général de la BnF, présentement dans un processus de renouvèlement, permet d'obtenir les références à la majorité des documents tenus par la BnF. Il ne contient cependant pas la plupart des manuscrits et des archives, les médailles et les antiques ainsi que les livres en écritures non latines, qui ont tous leur propre catalogue (BnF, 2014-b). Au cours des dernières années, la BnF a mis en place un grand nombre d'outils favorisant l'accès aux documents numérisés. D'abord, on note la bibliothèque virtuelle Gallica, accessible en ligne dès 1997 qui se bonifie chaque année d'un grand nombre de nouveaux documents numérisés. Avec plus de 3,2 millions de documents disponibles en 2014 (BnF, 2015-b), Gallica collabore avec plusieurs partenaires (bibliothèques et archives des collectivités territoriales, structures régionales de coopération, sociétés savantes, partenaires de l'enseignement supérieur et de recherche et autres) afin d'assurer une offre globale. Ainsi, les documents du domaine public consultables directement sur l'interface de Gallica proviennent des collections patrimoniales de

la BnF, des collections des partenaires numérisés par la BnF et les collections des partenaires numérisés par ces derniers, pour un total de 90 % des documents. Le 10 % restant est consultable sur des sites externes et il s'agit de collections de bibliothèques partenaires moissonnées à l'aide de la norme OAI-PMH (BnF, 2013). Cette norme permet d'échanger des métadonnées grâce à des entrepôts interrogeables. Les fournisseurs de données donnent donc accès à leur catalogue et les « moissonneurs » peuvent formuler des requêtes. Les résultats sont fournis en format XML et incorporent souvent du Dublin Core. Gallica propose plusieurs types de documents, soit des imprimés numérisés, des manuscrits, des documents sonores, des documents iconographiques, des cartes et des plans. Sur la page d'accueil du site officiel, on présente le nombre de documents par format (voir figure 13). Les documents numérisés sont sélectionnés dans l'objectif de « […] constituer une bibliothèque encyclopédique et raisonnée, représentative des grands auteurs français et des courants de recherche et de réflexion par-delà les siècles » (BnF, 2015-c). Rendre accessibles de tels documents joue un rôle dans le succès de Gallica, dans la mesure où il s'agit d'une bonne occasion de diffuser des documents qui ne peuvent être manipulés dû à leur rareté ou leur fragilité, et ce, à travers le monde. De plus, Europeana moissonne régulièrement le contenu de Gallica grâce à une collaboration entre les deux instances, ce qui permet une plus grande diffusion de ce contenu.

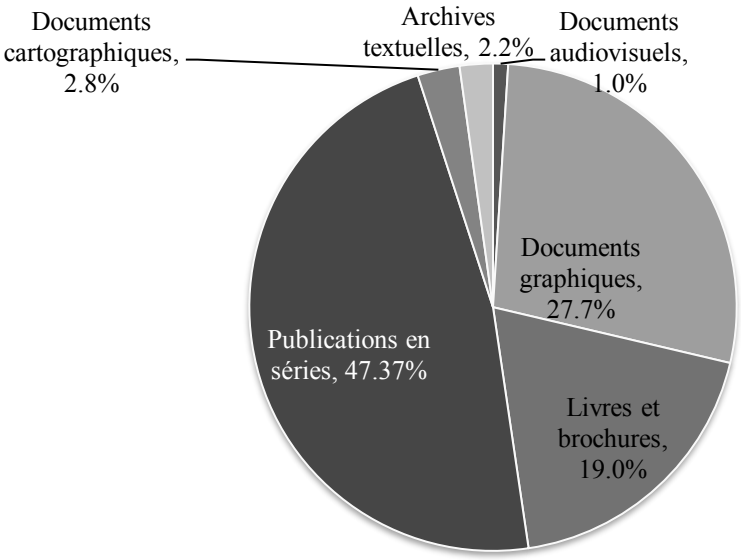


Figure 13. Répartition du nombre de documents disponibles sur Gallica en fonction du type de document en date du 26 avril 2016

Parallèlement à Gallica, la BnF a mis sur pied deux autres interfaces proposant du contenu numérique, soit la Banque d'images du département de la Reproduction ainsi que Mandragore (la base des manuscrits enluminés). Dans le premier cas, il s'agit d'un service de numérisation et d'impression qui permet à un usager de sélectionner un produit et de payer un tarif donné pour l'obtenir. En ce qui a trait à Mandragore, on y retrouve 170 000 notices indexées grâce à un vocabulaire de plus de 18 000 descripteurs, dont 80 000 notices contiennent une image numérisée (BnF, s.d.). Cette base de données permet des recherches avancées ainsi que des recherches par thématique.

Enfin, le projet data.bnf.fr, qui sera abordé en détail plus loin, est en place depuis 2011 et regroupe les données provenant de plusieurs catalogues de la BnF, soit le catalogue général, le catalogue des archives et des manuscrits et Gallica.

4.2.4 Diffusion et collaboration

La BnF est très active tant au point de vue international que national et produit plusieurs projets d'innovation numérique pour diffuser son contenu. Son implication dans des projets suivant les standards du Web de données démontre une volonté d'évolution de sa façon de faire afin de trouver de nouvelles voies pour améliorer la diffusion et le partage de données.

4.2.4.1 Médias sociaux

Au niveau de la diffusion, la BnF est aussi présente sur les réseaux sociaux. En date du 28 novembre 2015, on constate 67 062 mentions « J'aime » sur sa page Facebook, 143 000 abonnés à son fil Twitter et un compte LinkedIn sur lequel on retrouve 4 318 abonnés. De plus, sur Facebook, la présence est aussi assurée grâce à d'autres pages supplémentaires, c'est-à-dire des pages dédiées à Gallica, Classes BnF, Bibliothèque numérique des enFants, Arlequin – BnF Arts du spectacle et au Centre national de la littérature jeunesse. Sur Twitter, on retrouve Gallica BnF, Labo BnF, Dépôt légal du Web, service presse BnF et BnF monde (dédié à l'actualité

internationale). Très active sur ces réseaux, la BnF s'assure donc d'être en constante communication avec ses usagers.

4.2.4.2 Collaboration nationale

La BnF se démarque entre autres par sa collaboration active avec les bibliothèques françaises. Dans l'objectif d'augmenter la diffusion et de valoriser les documents patrimoniaux, elle a mis sur pied le Réseau national de coopération, qui vise à apporter un soutien financier (le budget est de 2,49 millions d'euros pour l'ensemble des actions menées) ainsi qu'un service de formation et de conseils à l'ensemble des bibliothèques de la France. Le réseau est composé des pôles associés de Dépôt légal, des pôles associés documentaires, des partenaires de coopération documentaire et il réunit plus de 200 partenaires BnF (BnF, 2015-d). Les deux objectifs premiers du réseau sont la valorisation du patrimoine écrit par l'enrichissement du Catalogue collectif de France et la coordination de l'effort de numérisation (BnF, 2015-d). Ainsi, tel que mentionné précédemment, le programme de numérisation concertée permet d'enrichir la bibliothèque numérique Gallica grâce aux documents provenant des bibliothèques françaises qui souhaitent participer. Les enjeux liés à cette initiative de numérisation à l'échelle nationale sont triples (BnF, 2015-e) :

1. Permettre un accès facile au patrimoine national, régional et local ;
2. Assurer une meilleure visibilité des documents patrimoniaux français sur Internet ;
3. Mettre de nouvelles sources à la disposition des chercheurs.

Grâce à ces deux services en particulier, on constate que la BnF joue un rôle non seulement en ce qui a trait à la numérisation et à la mise en ligne de documents patrimoniaux, mais aussi au niveau de l'éducation et de la formation des professionnels de l'information. Des bibliothèques moins importantes qui souhaitent donc diffuser leurs ressources ont donc la possibilité de le faire grâce au soutien de la BnF.

4.3 Comparaison entre les deux institutions

Cette section permet de jeter un regard d'ensemble sur les deux institutions en comparant les deux organisations ainsi que le travail de diffusion et de collaboration commun effectué par chacune.

4.3.1 Organisation

Il est difficile de comparer les effectifs de BAnQ et de la BnF dans la mesure où le processus de recrutement diffère largement d'une institution à l'autre et la fonction publique française présente d'énormes différences de fonctionnement par rapport à celle du Québec. La BnF divise ses différents métiers entre cinq catégories distinctes, soit :

1. Services au public, traitement des collections;
2. Médiation culturelle, communication;
3. Administration, gestion;
4. Technique, logistique et prévention;
5. Encadrement.

Il n'a pas été possible d'avoir accès au nombre exact d'emplois par catégories, mais plutôt uniquement aux pourcentages (BnF, 2014-c). Le tableau 6 permet donc de déduire le nombre d'emplois par catégories (à partir du total de 2 393 employés pour la BnF cité précédemment).

Tableau 6. Approximation du nombre d'emplois par catégories de métier en fonction du pourcentage à la BnF en 2014

Catégories de métiers	**Pourcentage (%)**	**Nombre d'emplois**
Services au public, traitement des collections	60	1436
Médiation culturelle, communication	8	192
Administration, gestion	12	287
Technique, logistique et prévention	11	263
Encadrement	9	215
Total	100	2393

Alors que BAnQ présente plutôt ses emplois en trois catégories distinctes, soit le personnel d'encadrement, les professionnels et conseillers en ressources humaines et les employés de soutien. On peut donc reclasser les catégories de métiers proposés par la BnF et les inclure dans ces trois catégories de BAnQ (en répétant le nombre d'employés) :

- Personnel d'encadrement (BAnQ) (43)
    - Encadrement (BnF) (215)
- Professionnels et conseillers en ressources humaines (BAnQ) (219)
    - Services au public, traitement des collections (BnF) (1436)
    - Médiation culturelle, communication (BnF) (192)
- Employés de soutien (BAnQ) (455)
    - Administration, gestion (BnF) (287)
    - Technique, logistique et prévention (BnF) (263)

Ainsi, le tableau 7 permet de comparer les effectifs entre la BnF et BAnQ, par catégories d'emplois.

Tableau 7. Répartition des emplois de BAnQ et de la BnF par catégories, en nombre d'emplois et en pourcentage

	Institutions			
	BAnQ		BnF	
Catégories d'emplois	Nombre d'emplois	Pourcentage (%)	Nombre d'emplois	Pourcentage (%)
Personnel d'encadrement	43	6	215	8,98
Professionnels et conseillers en ressources humaines	219	30,54	1628	68,04
Employés de soutien	455	63,46	550	22,98
Total	717	100	2393	100

On constate que le pourcentage des employés de soutien est trois fois plus élevé à BAnQ qu'à la BnF. Il est très important de considérer cependant le plus grand besoin d'employés travaillant avec le public et au traitement des collections à la BnF étant donné que la bibliothèque est beaucoup plus grande et dessert une population plus importante que BAnQ. Il est aussi intéressant de constater la différence au niveau du nombre total d'employés dans chacune des institutions. Cet aspect est pertinent, car on peut déduire que l'équipe étant responsable des technologies de l'information ainsi que de tout ce qui touche aux documents numériques est plus importante du côté de la BnF que du côté de BAnQ. Cette constatation est importante, car elle implique que dans la mesure où BAnQ souhaitait mettre en place un projet semblable à celui de la BnF, il lui faudrait prévoir une équipe dédiée importante par rapport aux ressources qu'elle possède déjà.

4.3.2 Diffusion et collaboration

Nous présenterons ici les projets pour lesquels BAnQ et la BnF sont tous les deux partenaires. En étant actives au sein de la communauté internationale, les deux institutions privilégient le partage et le travail collaboratif, principalement en ce qui a trait aux documents patrimoniaux. De plus, de telles initiatives permettent une meilleure diffusion des collections à travers le monde et respectent les engagements relatifs à leur mission respective.

4.3.2.1 Réseau francophone numérique (RFN)

BAnQ et la BnF sont des institutions membres, aux côtés de 23 autres institutions, du Réseau francophone numérique (RFN) dont la mission est triple (RFN, s.d.-a) :

1. Préserver le patrimoine patrimonial grâce à la numérisation et le diffuser au sein d'un large public;
2. Organiser des stages de formation et élaborer des outils didactiques afin de communiquer les connaissances aux institutions documentaires de la Francophonie;
3. Offrir un forum privilégiant les échanges autour des enjeux liés à l'ère numérique.

En 2014-2015, BAnQ a déposé 36 documents québécois sur le site du RFN (BAnQ, 2015). L'interface permet ainsi d'effectuer des recherches au sein des documents fournis par les institutions participantes grâce entre autres à la norme OAI-PMH. Le RFN propose aussi des documents de formation, entre autres sur les pratiques exemplaires quant à la numérisation de documents. Ces documents sont principalement fournis par BAnQ et la BnF. On y retrouve, entre autres, des présentations sur le nommage des fichiers numériques, les formats de numérisation et l'archivage de documents numériques.

4.3.2.2 Bibliothèque numérique mondiale

Les deux institutions participent aussi à la Bibliothèque numérique mondiale, une initiative de l'UNESCO et la Bibliothèque du Congrès, en ligne depuis 2009. Du côté de BAnQ, on retrouve 110 documents sur l'histoire et la géographie, les sciences sociales, les arts, la technologie et la religion provenant de ses collections (Bibliothèque numérique mondiale, s.d.-a). En 2014-2015, BAnQ y a déposé 30 nouveaux documents (BAnQ, 2015). Pour ce qui est de la BnF, elle y a déposé au total 133 documents (Bibliothèque numérique mondiale, s.d.-b). La Bibliothèque numérique mondiale propose ses propres éléments de description des métadonnées et accepte les formats MARC, Metadata Object Description Schema (MODS), Dublin Core et UNIMARC (Bibliothèque numérique mondiale, s.d.-c).

4.3.2.3 *Virtual International Authority File* (VIAF)

Depuis septembre 2015, BAnQ verse ses fichiers d'autorité vers le projet VIAF, un projet collaboratif auquel un grand nombre de bibliothèques nationales participent. Le projet est géré par OCLC et a comme objectif de « […] faire baisser le coût et de valoriser les fichiers d'autorité des bibliothèques par l'appariement et l'établissement de liens entre les fichiers d'autorité des bibliothèques nationales, et en rendant cette information disponible sur le web. » (VIAF, s.d.). Ce service est gratuit et accessible à tous. La BnF est aussi un partenaire très actif depuis 2007 et verse dans VIAF les données d'autorité sur les noms de personnes, mais aussi sur les collectivités, les noms géographiques, les œuvres et les expressions (BnF, 2015-f). Cet

aspect est particulièrement intéressant dans la mesure où VIAF est régulièrement mis à jour dans le cadre du Web de données, comme nous l'avons vu à la section 2.5.2 et nous y reviendrons également à la section 6.1.10.

4.3.2.4 WorldCat

Le service WorldCat est une base de données bibliographiques en ligne permettant d'effectuer une recherche fédérée au sein de catalogues provenant de milliers de bibliothèques. BAnQ et la BnF déposent leurs notices dans cette base de données, fournissant ainsi un meilleur accès à leurs collections ainsi qu'une meilleure visibilité.

4.3.2.5 Archives Canada-France

Aux côtés de Bibliothèque et Archives Canada (BAC), BAnQ et la Direction des Archives de France sont partenaires du projet Archives Canada-France, une base de données contenant plus d'un million d'images traitant de la Nouvelle-France. Présentement disponibles sur un site temporaire, les institutions travaillent à améliorer ce service et à mettre de l'avant cette exposition virtuelle (BAC, 2016).

5. Description des états d'avancement du projet de Web de données au sein des bibliothèques

Ce chapitre a comme objectif de présenter l'état d'avancement des projets liés au Web de données et au Web sémantique au sein de chacune des bibliothèques nationales. Il est évident que la BnF est nettement plus avancée quant à sa participation dans ce mouvement collaboratif, mais on constate tout de même la présence de petits projets d'exploration des technologies liées au Web sémantique du côté de BAnQ.

5.1 Bibliothèque nationale de France

La Bibliothèque nationale de France a su, au fil des cinq dernières années, se démarquer grâce à ses initiatives relatives au Web sémantique. En 2013, les projets data.bnf.fr et Gallica furent récipiendaires du *Stanford Prize for Innovation in Research Libraries* (SPIRL) dont l'objectif est de reconnaître et mettre de l'avant les projets innovateurs qui permettent aux usagers des bibliothèques d'avoir accès à de nouveaux services (Stanford University Libraries, s.d.). Toujours en évolution et en croissance, le projet data.bnf.fr fut mis sur pied afin de répondre à plusieurs objectifs (BnF, 2015-g) :

- assurer la visibilité des données de la BnF grâce à leur exposition sur le Web ;
- permettre la réutilisation de ces données par d'autres acteurs du domaine grâce à une politique de licence ouverte ;
- assembler les données de la BnF en un seul et même endroit ;
- participer à la coopération et à l'échange des métadonnées grâce à la création de liens entre des ressources structurées.

Ainsi, une moindre modification au sein d'une application de BnF (p. ex. Gallica, BnF archives et manuscrits, catalogue général) affecte directement le fonctionnement de data.bnf.fr. Pour l'usager moyen, cette plateforme permet d'être redirigé aux différents services offerts et de constater que des informations sur un sujet donné peuvent être retrouvées ailleurs que dans

le catalogue général, par exemple. Il s'agit véritablement d'un tremplin vers les autres plateformes et services. L'usager peut aussi passer d'un service de la BnF à un autre pour ensuite être redirigé à des ressources extérieures, soit d'autres jeux de données produits par d'autres organisations, traitant de l'auteur, de l'œuvre, du thème ou des lieux recherchés.

À partir des ressources que sont les collections numérisées (Gallica), BnF archives et manuscrits et BnF catalogue général, data.bnf.fr permet donc le regroupement et l'alignement des données qui en découlent pour ensuite les rendre visibles aux humains grâce à l'interface Web de data.bnf.fr. Ces données deviennent aussi traitables par les machines, grâce aux données structurées en RDF. Ces données sont ensuite exposées et peuvent être récupérées librement. Par exemple, dans le cadre du projet TELplus financé par la Commission européenne, toutes les notices d'autorité sujet du référentiel RAMEAU (Répertoire d'autorité-matière encyclopédique et alphabétique unifié) ont été converties en RDF SKOS (voir section 2.6.1). Ce référentiel est donc maintenant libre d'accès et peut être récupéré par tous (BnF, 2015-g). Dans les sections suivantes (5.1.1 à 5.1.3), nous aborderons le projet data.bnf.fr, le projet OpenCat ainsi que le Système de Préservation et d'Archivage Réparti (SPAR).

5.1.1 Le projet data.bnf.fr

Cette section a comme objectif de présenter les particularités relatives à la plateforme data.bnf.fr, les ressources mobilisées pour la mise en place d'un tel projet, les étapes nécessaires, les bénéfices pour la BnF et les prochains défis.

5.1.1.1 Exemple illustrant certaines fonctionnalités de la plateforme data.bnf.fr

L'exemple ci-dessous démontre qu'en faisant une recherche rapide sur data.bnf.fr en date du 18 mars 2016 pour le lieu « Montréal », on obtient les résultats classés par « Auteurs » (4 résultats), « Organisations » (517 résultats), « Œuvres » (5 résultats), « Thèmes » (46 résultats), « Lieux » (9 résultats), « Spectacles » (2 résultats) et « Périodiques » (83 résultats) (voir figure 14).

Après avoir choisi le lieu « Montréal (Québec, Canada) », l'usager est redirigé vers une page remplie d'informations (voir figure 15) telles que les coordonnées géographiques, une carte interactive, une présentation des documents disponibles, numérisés ou non, etc.

Figure 14. Capture d'écran des résultats obtenus suite à une recherche simple avec le mot-clé « Montréal » sur le site data.bnf.fr en date du 18 mars 2016

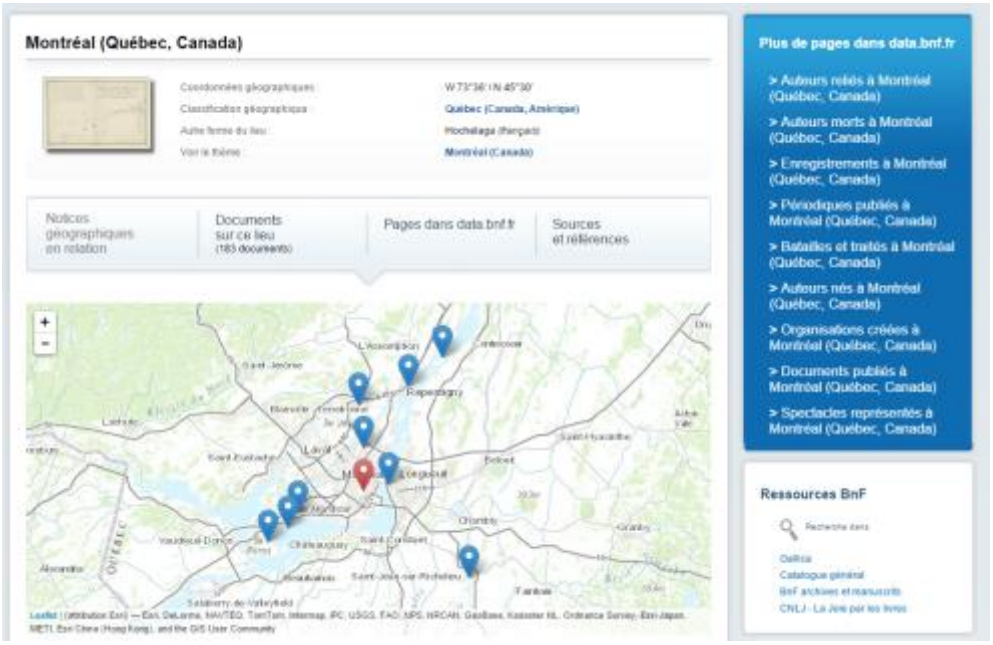


Figure 15. Extrait de la page dédiée à Montréal (Québec, Canada) sur le site data.bnf.fr

On y retrouve aussi des suggestions de pages liées à la recherche se trouvant sur data.bnf.fr tels que :

- Auteurs reliés à Montréal (Québec, Canada);
- Auteurs morts à Montréal (Québec, Canada);
- Batailles et traités à Montréal (Québec, Canada);
- Spectacles présentés à Montréal (Québec, Canada);
- etc.

Figure 16. Extrait de la page dédiée à Montréal (Québec, Canada) sur le site data.bnf.fr où l'on peut voir les liens vers les sites externes

On offre aussi la possibilité à l'usager d'être redirigé vers les ressources de la BnF, mais aussi vers des sites extérieurs tels que le Catalogue collectif de France, Europeana, OCLC WorldCat et Wikipedia (voir figure 16). Ce type d'exemple illustre particulièrement bien les possibilités et les occasions de rencontres non attendues, soit le concept de sérendipité mentionné préalablement (section 3.2.5). Ainsi, pour le plus de besoins potentiels possibles des usagers, des pages sont créées pour des lieux ou des années par exemple, et celles-ci sont ensuite reliées à d'autres pages afin d'augmenter leur potentiel de consultation et de visualisation. Aussi, l'outil permet de faire une seule recherche fédérée, donc facilite grandement le processus de repérage d'information pour l'usager. Notons que l'usager a la possibilité d'interroger un point d'accès SPARQL afin d'accéder aux données ou de télécharger les données via des *dumps*.

5.1.1.2 L'importance du modèle FRBR et la « FRBRisation » des catalogues

Pour évoluer vers son projet de Web de données, la BnF a dû procéder à une « FRBRisation » de son catalogue, c'est-à-dire le fait d'y appliquer le modèle conceptuel *Functional Requirements for Bibliographic Records* (FRBR). On retrouve, auprès de ce modèle, deux extensions complémentaires : le modèle *Functional Requirements for Authority Data* (FRAD) et le modèle *Functional Requirements for Subject Authority Data* (FRSAD). En 1990, le programme *IFLA Universal Bibliographic Control and International MARC* (UBCIM) ainsi que l'*IFLA Division of Bibliographic Control* ont mis sur pied un séminaire visant à examiner la raison d'être et la nature des notices bibliographiques (Howarth, 2012; Willer et Dunsire, 2013). Les participants au séminaire ont reconnu que l'environnement dans lequel les principes et standards liés au catalogage évoluaient avait changé drastiquement attribuable à la présence de systèmes de plus en plus automatisés et aux catalogues partagés (IFLA Study Group on the Functional Requirements for Bibliographic Records, 1998).

À la base, le modèle a été publié comme un modèle « entité-association » (*entity-relationship*), un modèle de données visant à distinguer les objets et leurs associations, ce qui a permis de porter une attention particulière sur les besoins de l'usager (Callewaert, 2013; Willer et Dunsire, 2013). Ainsi, le modèle FRBR a été réfléchi de façon à créer un cadre qui permet d'identifier et définir les entités pertinentes pour l'usager, ce qui caractérise ces entités et les relations qui les unissent. Les besoins de l'usager sont donc les suivants : trouver une entité, identifier une entité, sélectionner une entité et obtenir un document. Les entités sont séparées en trois groupes distincts (IFLA Study Group on the Functional Requirements for Bibliographic Records, 1998). D'abord, le groupe 1 comprend les entités qui sont le produit d'une activité intellectuelle ou artistique, c'est-à-dire l'œuvre (création intellectuelle ou artistique donnée), l'expression (la réalisation intellectuelle ou artistique d'une œuvre), la manifestation (la matérialisation de l'une des expressions d'une œuvre) et l'item (un exemplaire donné d'une manifestation). Ce modèle est intéressant dans la mesure où il présente une distinction claire entre l'information et sa concrétisation matérielle. La figure 17 illustre les relations entre les

entités du groupe 1 (IFLA Study Group on the Functional Requirements for Bibliographic Records, 1998).

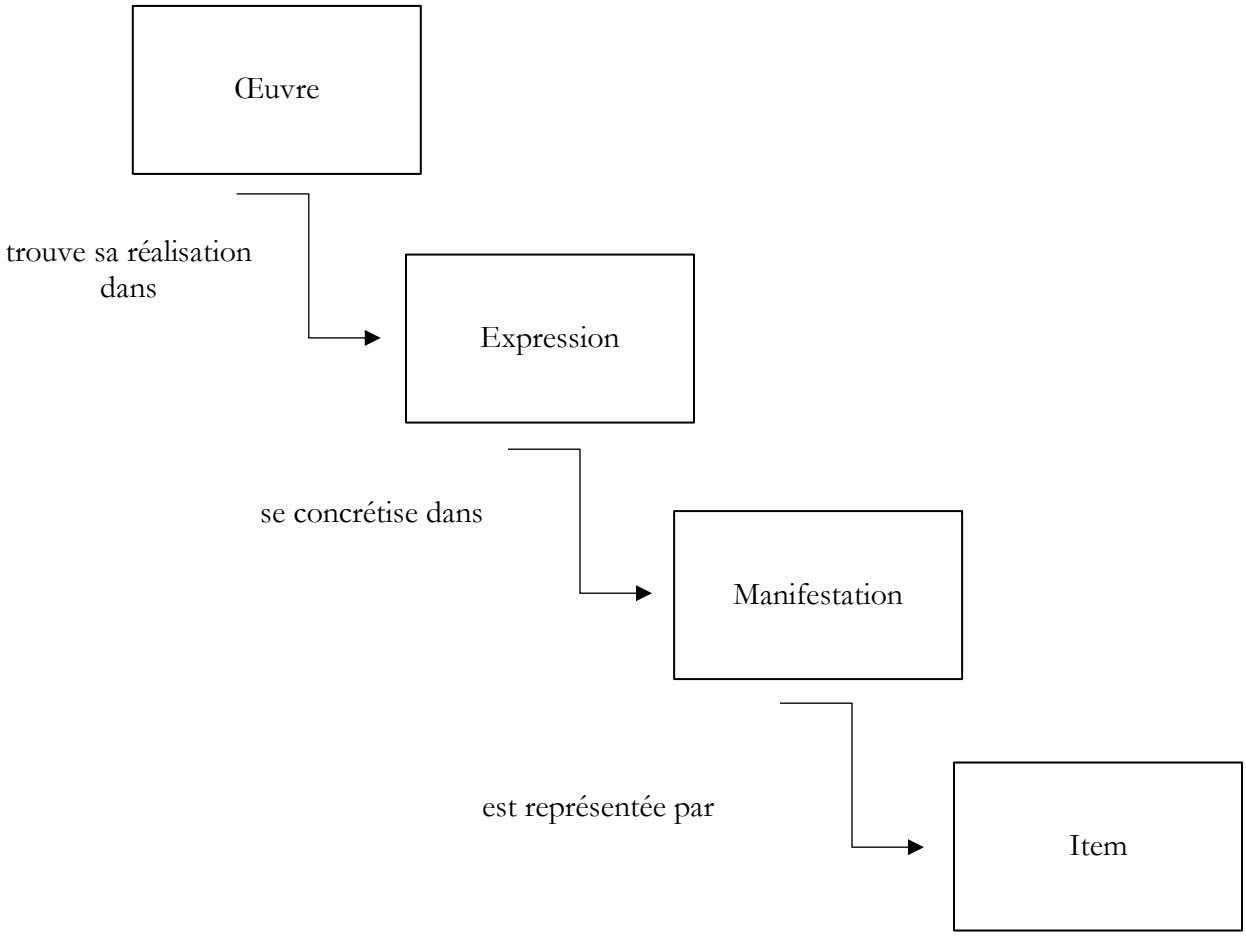


Figure 17. Entités du groupe 1 du modèle FRBR et les relations qui les unissent

Ensuite, le groupe 2 pour sa part est composé des entités qui ont la responsabilité du contenu, de la production matérielle, de la distribution et des droits relatifs aux entités du groupe 1. Ces entités sont les personnes et les collectivités. La figure 18 illustre les entités du groupe 2 et les liens de responsabilité qui les unissent au groupe 1 (IFLA Study Group on the Functional Requirements for Bibliographic Records, 1998). Les liens de responsabilité permettent donc de comprendre que les entités du groupe 1 peuvent être créées, produites, distribuées et propriétés d'une ou plusieurs entités du groupe 2.

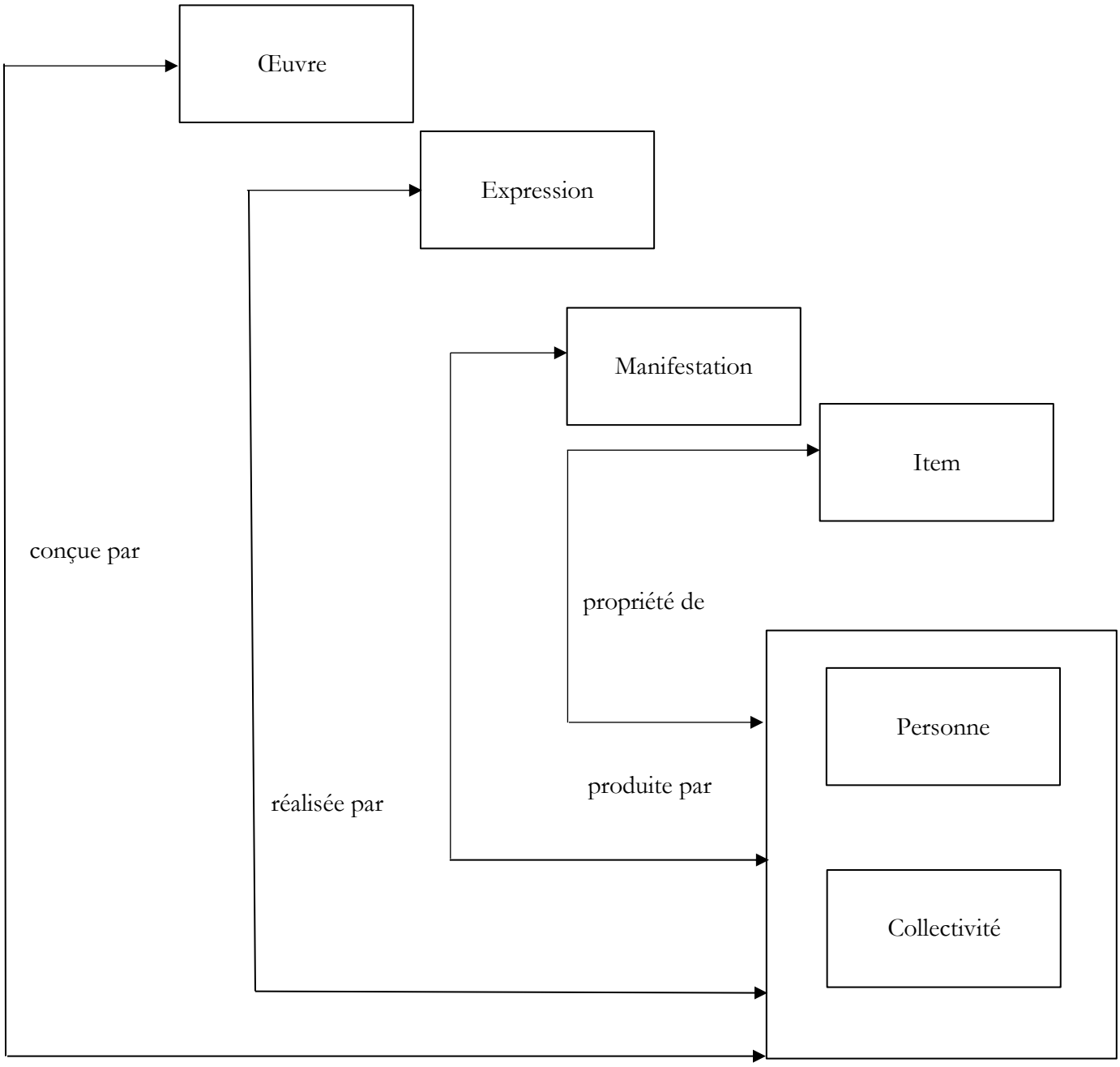


Figure 18. Liens de responsabilité entre les entités des groupes 1 et 2 du modèle FRBR

Finalement, les entités du groupe 3 représentent les sujets des œuvres, soit le concept, l'objet, l'événement et le lieu. Les modèles FRSAD et FRAD, pour leur part, sont venus bonifier plus tard le modèle FRBR en étendant les entités des groupes 2 et 3. L'objectif serait d'éventuellement proposer un modèle qui réunirait les trois composantes en son sein (Howarth, 2012). Les dernières années ont aussi vu une harmonisation du modèle FRBR au modèle sémantique de référence CIDOC CRM. CIDOC CRM est une ontologie visant la description et la représentation des collections de patrimoines culturels. Cette harmonisation a donné naissance à FRBRoo (FRBR orienté-objet) qui est un bon exemple de travail collaboratif misant sur les forces des deux modèles. Ainsi, on peut dire qu'il existe aujourd'hui deux modèles FRBR, soit FRBRer (FRBR *entity-relation*) et FRBRoo (Howarth, 2012; Willer et Dunsire, 2013).

Le modèle FRBR est au centre du mouvement d'évolution des catalogues. Le fait de libérer les données de notices bibliographiques peut permettre la création de nouvelles sortes de notices et le partage de ces mêmes données (Howarth, 2012). De plus, Bennet, Lavoie et O'Neill (2003) soulignent le fait que l'implantation du modèle FRBR dans les catalogues de bibliothèques permettrait d'accommoder un plus grand nombre de besoins provenant des usagers en présentant de nouvelles visions de la base de données bibliographique, d'augmenter la récupération efficace de documents grâce à une représentation hiérarchique des entités bibliographiques et, finalement, d'augmenter la productivité en ce qui a trait au processus de catalogage, en fusionnant l'information provenant de différentes notices. Le fait de présenter l'information bibliographique selon les entités du modèle FRBR permet aussi de ne pas dupliquer directement le catalogue, mais de permettre une organisation qui se rapproche plus du sens que donne l'usager au patrimoine culturel. De plus, dans le cadre du Web de données, la distinction faite entre « Œuvre », « Expression » et « Manifestation » permet de faciliter les liens entre les ressources étant donné que les autres communautés pourront réutiliser soit le graphe RDF complet qui décrit une publication donnée soit uniquement l'information liée à l'entité qui les intéresse (Le Bœuf, 2013). On constate aussi qu'une telle modification au niveau des catalogues permettra une granularité plus fine de l'information bibliographique, car on passera de la description par notices à la description par données. En mariant le modèle

conceptuel FRBR aux règles de catalogage RDA et à la syntaxe RDF, on permet donc une nouvelle présentation de l'information bibliographique en répondant aux standards du Web de données et, ainsi, une meilleure visibilité sur le Web.

5.1.1.3 Ressources mobilisées

Depuis le début du projet, une équipe de six personnes a été embauchée à temps plein pour travailler sur le projet data.bnf.fr. Cet effectif devrait être maintenu au fur et à mesure que le projet évolue. Il nous a été impossible d'obtenir le coût du total du projet, mais nous avons eu l'information selon laquelle la somme déboursée tournerait autour de quelques centaines de milliers d'euros. En ce qui a trait aux ressources informatiques, huit serveurs sont mobilisés pour héberger les données de data.bnf.fr et deux plateformes sont utilisées pour tester les développements.

5.1.1.4 Étapes relatives à la mise sur pied du projet

Une première phase de réflexion a d'abord eu lieu afin de permettre une réflexion sur la refonte et l'évolution du catalogue vers le modèle FRBR dans l'objectif d'assurer son optimisation dans le cadre du Web sémantique. Puis, une première étude de préfiguration a été faite avec un premier prestataire, la compagnie Capgemini. Par la suite, deux contrats ont été signés avec un nouveau prestataire, Logilab, qui s'apprête à débuter un troisième contrat, correspondant à une autre phase de réalisation, qui devrait être d'une durée de trois ans, à partir de septembre 2016. Ces phases de réalisation visent à mettre sur pied les prochaines phases du projet suite à une évaluation des besoins.

Les étapes de mise sur pied du projet, décrites ci-dessous, sont les suivantes :

- Transformation des données bibliographiques ;
- Enrichissement des données bibliographiques ;
- Création de liens vers des sources internes et externes ;

- Insertion des jeux de données dans les pages HTML et création des *dumps* RDF ;
- Création des données structurées internes.

Afin que le tout soit fonctionnel, il est nécessaire de transformer les données bibliographiques provenant de différents jeux de données, de les enrichir et de les relier à des ressources internes et externes. Pour sa part, data.bnf.fr offre aussi la possibilité de visualiser les pages HTML, donc il est nécessaire d'insérer ces jeux de données dans ces pages pour qu'elles soient comprises par l'humain. Le processus nécessite ainsi la transformation des formats MARC et des informations en format *Encoded Archival Description* (EAD) grâce à des techniques de conversion des données en RDF (BnF, s.d-b). Ce processus de modélisation a un impact majeur sur l'enrichissement et l'alignement des données et le tout est basé sur les entités de la modélisation conceptuelle qu'est FRBR, c'est-à-dire les groupes 1 (ici, œuvre uniquement, pour le moment[26]), 2 (les personnes ou les collectivités) et 3 (sujet, concept, lieu, événement, objet) (BnF, 2015-g; Simon, Di Mascio, Michel et Peyrard, 2014; Simon, Wenz, Michel et Di Mascio, 2013). Le processus d'alignement des données est le fait de mettre en correspondance des concepts afin de créer des liens entre les ressources internes et externes. Il s'agit ainsi de liens entre des entités qui sont équivalentes, similaires ou connexes au niveau du sens. Ces liens se font à l'aide de vocabulaires, d'ontologies, de schémas ou de jeux de données externes (Baker et al., 2011).

Ensuite, les données sont intégrées dans une architecture qui permet la création à la fois des pages HTML et des jeux de données en RDF (BnF, s.d-b). Une fois toutes les données alignées, il est ensuite possible pour un usager d'effectuer une recherche visant à trouver des documents directement via l'interface pour un concept, un auteur ou même un lieu. Les liens vers les ressources externes se font grâce à un outil sous licence libre (Nazca), dont le langage de programmation est Python et développé par la société française Logilab sur laquelle nous reviendrons sous peu. Finalement, des données structurées internes de Schema.org ainsi que

[26] Dans les prochaines phases de réalisation, la BnF souhaite procéder à l'application de ces techniques aux autres entités du groupe 1 du modèle FRBR.

Open Graph Protocol ont été intégrées dans les pages HTML afin d'aider les moteurs de recherche et les médias sociaux à trouver, en quelque sorte, les pages Web de data.bnf.fr (Simon, Di Mascio, Michel et Peyrard, 2014) (voir section 2.7). Par exemple, la description d'une page, son titre, son URL et l'image qui s'y rattache seront encodés dans le code HTML selon le protocole *Open Graph Protocol*[27], créé par Facebook mais maintenant géré par l'*Open Web Foundation*. Ainsi, cet ajout permettra à la page Web de donner aux médias sociaux de l'information par rapport à celle-ci. Ces données sont aussi prises en compte par les moteurs de recherche.

Logilab est le principal partenaire de data.bnf.fr dans la mesure où il travaille en collaboration avec la BnF depuis le début du projet. Au moment de le mettre sur pied, la décision fut prise de créer une base de données relationnelle utilisant le langage SQL, plutôt que d'aller vers un *triplestore*. Les *triplestores* sont des bases de données conçues pour le stockage et la récupération de données RDF – on y retrouve donc uniquement des triplets. Selon Simon, Di Mascio, Michel et Peyrard (2014), cette décision est basée à la fois sur le prix moindre, mais aussi sur leur volonté de baser un site Internet reposant sur de nouvelles technologies du Web sémantique sur un outil déjà connu et maîtrisé. Ainsi, il serait plus facile pour tous les acteurs (bibliothécaires et informaticiens) participant au projet de près ou de loin de comprendre les enjeux, mais aussi d'interagir les uns avec les autres. Cette base de données relationnelle est ainsi une sorte de pivot où l'on retrouve toutes les données provenant des autres applications en différents formats, soit RDF, JSON, CSV et HTML. Les données sont ensuite adaptées aux ontologies qui peuvent être modifiées par la suite si nécessaire.

5.1.1.5 Bénéfices

Les bénéfices du projet data.bnf.fr pour la BnF ont été nombreux. D'abord, on a constaté un nombre important de visites, soit maintenant 13 000 visiteurs uniques par jour, dont 81 % proviennent des moteurs de recherche. Puis, un plus grand nombre de visites d'usagers redirigés

[27] http://ogp.me/

(à partir de data.bnf.fr) a aussi été recensé sur les différentes applications de la BnF (Catalogue Général, BnF Archives et Manuscrits, Gallica, etc.), étant donné que les usagers, suite à une recherche sur la plateforme data.bnf.fr, ont tendance à aller consulter les différents documents au sein des autres services offerts.

Ensuite, plusieurs réutilisations des données ont été faites, grâce au point d'accès SPARQL. L'un des réutilisateurs les plus importants est Wikidata. Il s'agit d'un projet de la Wikimedia Foundation qui se définit comme une « (…) base de données libre, collaborative et multilingue, qui collecte des données structurées pour alimenter Wikipédia, Wikimedia Commons, les autres projets et bien plus encore. » (Wikidata, 2015). On constate aussi l'utilisation des données dans le cadre de projets de recherche effectués dans le cadre de fouille de données et de documents[28]. Ces pratiques de réutilisation sont intéressantes dans la mesure où elles répondent à l'objectif visant à encourager la création de nouvelles applications et nouveaux services grâce à l'accès aux données, mais aussi parce qu'elles permettent à la BnF de renforcer ses liens avec d'autres entités qui produisent et/ou réutilisent des données liées.

Finalement, un autre bénéfice tangible est le fait que le code de data.bnf.fr a pu être réutilisé pour la mise sur pied d'autres projets internes, tel que le projet Registre des Livres Indisponibles en Réédition Électronique (ReLIRE)[29], dont l'objectif est de pouvoir repérer toutes les œuvres indisponibles du XXe siècle. Une œuvre indisponible, selon la loi française, est une œuvre toujours sous droit d'auteur, qui a été publiée en France entre le 1er janvier 1901 et le 31 décembre 2000 et qui ne fait plus l'objet d'une diffusion commerciale ou qui n'est plus publiée sous forme imprimée ou numérique (BnF, 2014-d). Ce registre est donc un outil pour les auteurs, les détenteurs de droits et les éditeurs qui souhaitent vérifier si leur titre est présent dans la liste et d'exercer leurs droits si nécessaire. Ainsi, il était nécessaire pour la BnF de repérer les auteurs qui étaient présents en double dans les catalogues afin d'éviter les erreurs et d'améliorer la qualité de ceux-ci. La BnF nous indiquait qu'ils estiment qu'il y en aurait près d'un million. Pour ce faire, les techniques de partitionnement des données (*clustering*)

[28] Par exemple : http://resultats.hypotheses.org/518

[29] https://relire.bnf.fr

permettant de procéder à la « FRBRisation » dans le cadre de data.bnf.fr ont été réutilisées pour le projet ReLIRE.

5.1.1.6 Prochains défis

On réfléchit présentement au sein de la BnF à de nouveaux objectifs à atteindre pour le projet ainsi qu'aux axes stratégiques qu'il prendra dans les prochaines années. On présente actuellement data.bnf.fr comme une interface d'innovation autour des catalogues en ligne et l'objectif serait maintenant de rendre l'interface plus lisible grâce à une homogénéisation des données. Cependant, cela implique encore plus de partitionnement des données et de « FRBRisation » du catalogue. Ainsi, on pourrait par exemple enrichir encore plus les données grâce à un apport externe, proposer une recherche plus exploratoire afin de permettre plus de sérendipité et développer des visualisations pour encourager la découverte et la réutilisation des données. Le travail à effectuer est donc au niveau de l'expérience usager.

5.1.2 Le projet OpenCat[30]

Une fois data.bnf.fr stabilisé, la BnF s'est penchée sur la question à savoir si ses données ouvertes pourraient être utiles au sein des autres bibliothèques. L'usager étant au centre des préoccupations, on constate un effort particulier en ce qui a trait à la fiabilité des données proposées, mais aussi dans la tentative d'amélioration de l'expérience de recherche en soi. Les volontés de partager les données avec le plus grand nombre, de faciliter la recherche pour les usagers grâce à la modélisation FRBR, de rendre l'expérience de recherche interactive par la présence de documents numérisés et de proposer aux usagers une seule liste de résultats provenant de différentes sources sont au centre du projet OpenCat. Le prototype OpenCat est un pivot pour des ressources différentes, mais complémentaires. Il est caractérisé par un regroupement d'œuvres, les données de la bibliothèque municipale de Fresne, des compléments

[30] Pour avoir accès au démonstrateur : https://demo.cubicweb.org/opencatfresnes/

bibliographiques, des informations contextuelles non-bibliographiques et des liens directs vers des ressources en ligne (BnF, 2013-c).

Le prototype est basé sur les mêmes règles régissant data.bnf.fr, c'est-à-dire, entre autres, le regroupement selon les entités proposées par le modèle FRBR et des liens vers des ressources externes. Ainsi, en collaboration avec la bibliothèque municipale de Fresnes (environ 4 000 œuvres), les données de son catalogue ainsi que celles de data.bnf.fr ont été versées vers le prototype (BnF, 2013-c). Les données d'une nouvelle bibliothèque permettent donc d'enrichir les résultats de recherche et illustrent bien une application du Web de données. Pour une petite institution, cette innovation est intéressante dans la mesure où cela permet d'augmenter la visibilité de la collection sur le Web et d'enrichir son offre de services. Le démonstrateur, pour sa part, permet aux bibliothèques de visualiser leurs collections liées aux autres ressources et démontre le potentiel que proposait OpenCat. Un tel outil est intéressant, car il permet de présenter concrètement des applications du Web sémantique au sein d'institutions documentaires et d'envisager les possibilités qui s'offriraient à eux s'ils souhaitaient participer à un tel projet en ouvrant leurs données et en les rendant accessibles.

5.1.3 Le Système de Préservation et d'Archivage Réparti (SPAR)

Il est aussi nécessaire de présenter rapidement l'outil mis en place par la BNF afin d'assurer la préservation numérique, soit le Système de Préservation et d'Archivage Réparti (SPAR). Cet outil est intéressant dans le cadre du Web de données, car il accorde une attention particulière à la gestion des métadonnées. Ainsi, data.bnf.fr et SPAR se doivent d'être complémentaires.

L'objectif de SPAR n'est pas limité au stockage sécurisé. En effet, étant basé sur la norme *Open Archival Information System* (OAIS), il permet d'assurer la préservation pérenne des documents numériques et d'assurer leur lisibilité ainsi que leur possibilité de réutilisation dans le temps, et ce, même si l'environnement technologique vient à changer et évoluer au fil du temps (Bermès et Fauduet, 2011; Caron, Ledoux, Reecht et Tramoni, 2015). On n'a qu'à

penser à l'évolution des différents formats de fichiers, propriétaires ou non, qui ont fait en sorte dans les dernières années que plusieurs documents sont devenus illisibles, ou encore à la disparition de certains logiciels ou matériels informatiques qui permettaient la consultation de types de documents donnés.

La norme OAIS, issue du domaine aérospatial, est en fait principalement un cadre de référence qui permet d'appliquer les différents concepts pertinents à la préservation de documents à long terme. Elle ne présente pas de marche à suivre en tant que telle et est plutôt une base pour développer par la suite des normes complémentaires. La norme est constituée de définitions de concepts et de responsabilités découlant de la mise sur pied d'une archive OAIS, de la description de deux modèles détaillés (modèle fonctionnel et modèle d'information), des perspectives pour la préservation ainsi que des critères quant à l'interopérabilité des archives. Elle ouvre ainsi la porte à des applications flexibles et créatives tout en assurant le respect des règles de base liées à la préservation de documents numériques ou non. La norme se base entre autres sur la notion d'« ensemble d'information », c'est-à-dire tout ce qui est traité dans l'archive, soit ce qu'on y verse, ce qu'on y décrit et préserve et ce qu'on diffuse aux usagers.

Ainsi, les ensembles d'information versés dans l'archive se nomment les SIP (*Submission Information Package*), les ensembles archivés sont nommés les AIP (*Archival Information Package*) et, finalement, les ensembles d'information diffusés aux usagers lors d'une requête sont les DIP (*Dissemination Information Package*). Afin d'assurer la traçabilité des documents une fois l'archivage effectué, il est aussi nécessaire de les décrire adéquatement en indiquant leurs caractéristiques techniques, leurs identifiants, leurs contenus intellectuels, leurs structures (particulièrement lorsque les documents sont composés de plusieurs fichiers), leurs contextes et raisons d'être ainsi que l'historique des opérations effectuées sur ceux-ci (2013-b).

La BnF utilise donc les métadonnées de préservation pour décrire les ensembles d'informations, soit METS (*Metadata Encoding and Transformation Standard*) et PREMIS (*Preservation Metadata : Implementation Strategies*). D'un côté, METS est un schéma XML

conforme à la norme OAIS utilisé comme un conteneur pour ces descriptions, c'est-à-dire qu'il permet de décrire les métadonnées descriptives, administratives et de structure d'un document numérique afin de faciliter la diffusion et le partage (Cantara, 2005). On retrouve donc, pour chaque document numérique, un fichier METS dans lequel sont décrites toutes les métadonnées qui s'y rattachent. Pour sa part, PREMIS est un regroupement de métadonnées (*PREMIS Data Dictionary*) visant à donner des informations techniques sur le document ainsi que sur son historique afin d'assurer de façon spécifique la préservation à long terme (Caplan, 2009). Finalement, BnF fait appel aux métadonnées du Dublin Core pour les informations de contenu intellectuel et aux identifiants ARK (identifiants pérennes) des notices bibliographiques pour assurer que la description soit la plus complète possible. On constate donc que les technologies utilisées du côté de data.bnf.fr et du côté de SPAR sont semblables et qu'une interopérabilité est nécessaire afin de garantir le bon fonctionnement du projet de Web de données et de la préservation pérenne des documents numériques.

On constate que la BnF a donc fait un travail important au niveau de son implication dans le Web de données et y travaille depuis un moment déjà. En se basant sur cette expérience, on remarque que le travail doit se faire de manière continue et implique la participation de plusieurs instances différentes. La BnF a aussi fait le choix d'utiliser cette nouvelle plateforme afin que les usagers puissent accéder aux ressources des différents catalogues en effectuant une seule recherche. On permet donc une meilleure accessibilité aux collections ainsi qu'une meilleure diffusion à grande échelle.

5.2 Bibliothèque et Archives nationales du Québec

Cette section a comme objectif de présenter les premiers efforts de BAnQ en ce qui a trait à la mise sur pied de projets faisant appel aux technologies du Web sémantique. L'institution n'en est présentement qu'à ses débuts dans ce domaine, malgré le travail déjà effectué depuis 2012 dans le cadre du projet du Réseau francophone numérique, que nous aborderons à la section 5.2.1. Notons aussi la publication de données ouvertes sur le portail gouvernemental depuis 2013. L'arrivée du Plan culturel numérique du Québec en 2014, présenté

par le ministère de la Culture et des Communications (MCC), semble offrir une occasion toute particulière à l'institution. En 2010, le MCC a commencé une campagne visant à consulter ses différentes clientèles afin de pouvoir dresser un portrait des efforts à faire dans le but d'assurer un virage numérique au Québec. De ces consultations est né le Plan culturel numérique du Québec, dont l'objectif est d'aider les milieux culturels à effectuer la transition vers le numérique pour que la province puisse rayonner sur les marchés locaux, nationaux et internationaux. Le projet s'articule autour de trois grands axes (Gouvernement du Québec, 2016-a) :

- Créer des contenus culturels numériques ;
- Innover pour s'adapter à la culture numérique ;
- Diffuser des contenus culturels numériques afin d'assurer leur accessibilité.

Ainsi, en chiffres, le Plan culturel numérique du Québec est constitué de plus de 50 mesures prises en charge par des organismes tels que BAnQ et d'un budget de 110 millions de dollars répartis sur sept ans. Dans ce cas-ci, la mesure qui nous intéresse est la suivante (mesure 06) : « Aider le réseau de la culture à s'approprier les technologies du Web sémantique afin de maximiser la présence des données culturelles québécoises dans le Web » (Gouvernement du Québec, 2016-b). Cette mesure a deux objectifs qui se divisent en plusieurs actions. D'abord, on demande la création d'un groupe d'experts dont les objectifs sont de coordonner et d'orienter les institutions culturelles du Québec dans la structuration et l'échange de données. Ensuite, la mesure 06 est aussi caractérisée par la mise sur pied d'un partenariat et des projets-pilotes avec le ministère de la Culture et de la Communication en France dans le but d'intégrer les données québécoises dans le Web sémantique francophone (Gouvernement du Québec, 2016-b).

Ce projet offre donc une occasion à BAnQ de pouvoir jouer un rôle important au sein de ce projet innovateur au Québec. Jusqu'à maintenant, peu de projets mettant en scène les technologies du Web sémantique ont été mis sur pied dans la province et le fait de prendre les rênes de cette mesure et d'en diriger les différentes étapes permettra peut-être de mettre en place les bases de plusieurs projets au sein des différentes institutions culturelles québécoises.

Nous présentons ci-dessous deux projets de BAnQ. D'abord, son implication dans le Réseau francophone numérique, puis BAnQ numérique, sa nouvelle bibliothèque numérique.

5.2.1 Initiative au sein du Réseau francophone numérique (RFN)

Certains concepts du Web de données ne sont cependant pas étrangers à BAnQ. Il y a quelques années, BAnQ a participé à un projet visant à rendre disponible l'ensemble des métadonnées des institutions membres du Réseau francophone numérique (voir section 4.3.2.1). Ces jeux de données peuvent être récupérés à partir du *triplestore* du RFN dans les syntaxes de sérialisation suivantes : RDF/XML, RDF/N-Triples, RDF/N3, RDF/Turtle et RDF/JSON (voir section 2.4.2). Afin que le téléchargement ne soit pas trop lourd pour les usagers, les jeux de données sont séparés par catégories, soit journaux, revues, livres, cartes et plans et audiovisuel. Un autre aspect important de ce projet est la mise sur pied d'un point d'accès SPARQL, qui donne la possibilité aux usagers d'effectuer des recherches directement sur l'interface du site dans ce langage de requête. Le point d'accès permet à la fois de faire des requêtes de recherche (SELECT) ainsi que des constructions de graphes de triplets (CONSTRUCT). Ensuite, ces données sont ouvertes et dépendent donc d'une licence d'utilisation, particulièrement dans ce cas-ci où plusieurs institutions sont concernées et déversent leurs métadonnées sur cette plateforme. Ainsi, les usagers sont en mesure « […] d'utiliser, de reproduire, de traduire, de compiler et de communiquer au public par quelque moyen que ce soit, en tout ou en partie, les données ouvertes disponibles […] » (RFN, s.d.-b). RFN demande cependant que l'usager mentionne la source dans le cas où il réutiliserait les données.

5.2.2 BAnQ numérique

En ligne depuis 10 ans, le portail de BAnQ se doit d'être mis à jour et, pour ce faire, les dirigeants ont mis sur pied un projet à long terme nommé BAnQ numérique. Développé en lien avec le Plan culturel numérique du Québec, il repose sur huit principes directeurs qui ont comme objectifs de fédérer l'ensemble des ressources numériques du patrimoine documentaire en un seul lieu et ainsi, de faciliter l'accès à l'information, d'encourager la découverte et la recherche

ainsi que le partage de connaissances (BAnQ, 2015-b). Les principes directeurs, disponibles sur le site Internet en version bêta sont les suivants :

1. Il faut rassembler l'offre numérique de BAnQ en un tout cohérent ;
2. L'expérience utilisateur est au centre de tout ;
3. La recherche, la découverte et le partage sont les fondements de BAnQ numérique ;
4. L'ouverture tous azimuts est son fil conducteur ;
5. Le contenu et les services sont adaptés à toutes les plateformes et à tous les écrans ;
6. Le contenu mène naturellement à la création de services connexes ;
7. De nature évolutive, le contenu est évalué et amélioré en continu ;
8. Le projet se veut rassembleur et transversal.

Ici, le quatrième principe est le plus important. En effet, celui-ci démontre une volonté d'encourager la diffusion des contenus et des métadonnées qui les caractérisent, de moissonner des données de l'externe et d'ouvrir les collections au plus grand nombre. Grâce à cette volonté de modernisation des technologies et cette ouverture d'esprit en ce qui a trait au partage des métadonnées, BAnQ numérique a l'occasion de bonifier son offre en faisant appel au Web sémantique. Quelques projets sont déjà en cours tels que l'ajout de données structurées internes dans les pages Web décrivant des événements organisés par BAnQ en format JSON-LD.

5.2.3 Dépôt numérique fiable (DNF)

De son côté, BAnQ travaille depuis quelque temps à la mise sur pied d'un dépôt numérique fiable qui a comme objectif d'assurer la préservation des ressources numériques à long terme. Ce dépôt numérique fiable est l'équivalent ou presque du SPAR basé sur la norme OAIS présenté à la section 5.1.3. Dans la mesure où un projet de Web de données verrait le jour à BAnQ, le DNF peut jouer un rôle important dans la mesure où l'objectif serait de créer un système interopérable où tous les constituants du système pourraient assurer la stabilité et la fiabilité de celui-ci. À ce modèle, on peut envisager l'ajout d'un *triplestore* qui permettrait le stockage des métadonnées structurées converties en RDF. La figure 19 propose une visualisation de ce que sera le DNF à BAnQ. Comme on peut le constater, les principales composantes du DNF pensé par BAnQ sont les suivantes :

- Un espace de réception ;
    - Permet l'entrée des paquets d'information SIP préparés par les clients internes et externes.
- Un espace de normalisation automatisé ;
    - Vérifie l'intégrité du SIP et que les règles établies par BAnQ ont été respectées. Cet espace de normalisation gère les formats, ajoute les métadonnées nécessaires à la conservation des documents et préparent les DIP pour la diffusion et les AIP pour l'archivage.
- Un entrepôt de conservation OAIS ;
    - Est en charge de la préservation numérique.
- Un espace de gestion des métadonnées ;
    - Permet que les informations relatives aux métadonnées soient exactes et les mêmes d'un système à un autre.
- Un interface d'accès pour les employés ;
- Un entrepôt de diffusion OAIS ;
    - Gère les activités liées à la diffusion pour les usagers et pour d'autres institutions partenaires.
- Un *triplestore*.
    - Met en place les technologies du Web sémantique et stocke les triplets RDF.

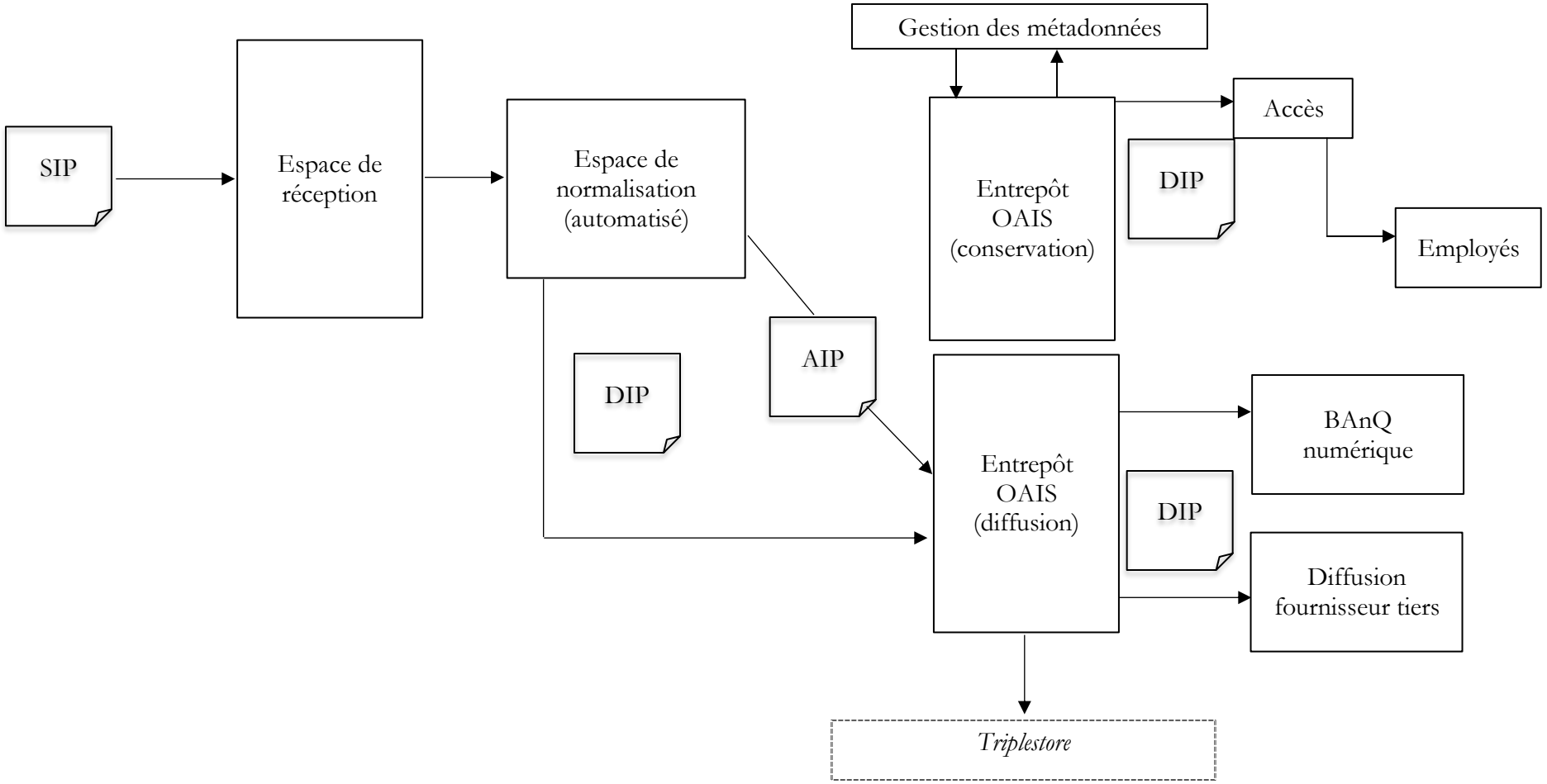


Figure 19. Modèle du DNF de BAnQ

On constate donc un intérêt de la part de BAnQ pour les technologies du Web sémantique, et ce, depuis la mise en place du site Web du RFN il y a quelques années. Cependant, un écart est remarqué entre le travail de la BnF et celui accompli à BAnQ. Le chapitre 6 vise entre autres à évaluer dans quelle mesure il serait possible pour BAnQ, en fonction des ressources qu'elle a, de mettre en place un projet de Web de données semblable au travail effectué par la BnF.

6. Étapes de mise sur pied d'un projet de Web de données et recommandations

Parmi la littérature disponible sur le Web de données et le Web sémantique, on retrouve quelques recommandations et méthodologies à suivre pour publier des jeux de données sur le Web (notamment Bizer et Heath, 2011; Van Hooland et Verborgh, 2014; W3C, 2014-b; Zengenene, Casarosa et Meghini, 2014). Cependant, il est difficile d'obtenir une version claire et condensée d'une marche à suivre afin d'appliquer ces technologies aux données bibliographiques. L'objectif de ce chapitre est de présenter les étapes nécessaires à la mise sur pied d'un projet de Web de données au sein d'une institution documentaire en fonction de leurs besoins et de la situation d'informatisation évolutive à laquelle ils doivent faire face. Nous porterons une attention particulière aux décisions prises par la BnF comme approches applicables et proposerons des possibilités pour BAnQ. Finalement, des recommandations pour BAnQ seront aussi formulées en vue d'une éventuelle implémentation.

6.1 Étapes de mise sur pied d'un projet de Web de données

Il est d'abord primordial, afin de favoriser l'interopérabilité, de rappeler les principes de base du Web sémantique présentés à la section 2.2, soit :

1. Nommer les ressources avec des URI pérennes ;
2. Utiliser des URI déréférençables (protocole HTTP) afin qu'il soit possible d'accéder à des informations sur les ressources ;
3. S'assurer que les URI déréférençables fournissent des informations pertinentes à l'aide des standards tels que RDF et SPARQL ;
4. Créer un réseau de liens avec d'autres URI provenant d'autres bases de données.

Les étapes nécessaires à la mise en place d'un projet de Web de données présentées ci-dessous sont les suivantes :

1. Comprendre la motivation et prise de conscience ;
2. Obtenir l'autorisation des parties prenantes ;

3. Établir une licence d'utilisation ;
4. Évaluer les compétences ;
5. Évaluer les jeux de données ;
6. Choisir le modèle de publication et évaluer les outils nécessaires ;
7. Attribuer les URI ;
8. Choisir son modèle de données et faire le *mapping* ;
9. Nettoyer les données ;
10. Enrichir les données en faisant des liens ;
11. Convertir les données en RDF ;
12. Évaluer les jeux de données ;
13. Publier les jeux de données.

Ces étapes sont le résultat d'une synthèse effectuée suite à la revue de la littérature. Les informations étant disséminées à travers différents textes, un travail d'analyse, de regroupement, de classement et de fédération de celles-ci a été effectué. Le but de l'exercice est de proposer un processus clair de publication dans le Web des données en se basant sur les standards préconisés par le W3C afin de faciliter l'appropriation des technologies du Web sémantique par les professionnels de l'information.

Les deux premières étapes (sections 6.1.1 et 6.1.2) ne sont pas spécifiques à un projet de Web de données et sont des exigences liées au facteur humain. En effet, elles pourraient s'appliquer à n'importe quelle gestion de projet d'implémentation nouvelle. Cependant, elles sont tout de même primordiales à son fonctionnement et son acceptation au sein de l'institution. En ce qui a trait aux approches retenues par la BnF et aux possibilités pour BAnQ pour ces deux étapes, nous n'avons malheureusement pas pu obtenir d'informations à ce sujet. Nous jugeons que les deux institutions, en leur qualité de bibliothèque nationale dépendant d'instances gouvernementales, effectuent ces deux étapes de manière régulière dans le cadre d'autres projets informatiques.

6.1.1 Comprendre la motivation et prise de conscience

Avant de commencer la mise en place d'un tel projet, il est nécessaire que tous les acteurs impliqués puissent avoir une vision globale de ce que signifient le Web de données et les technologies qui s'y rattachent, mais aussi les avantages et défis qui en découlent. L'importance réside dans la compréhension du rôle des sciences de l'information dans cette nouvelle extension du Web, dans l'amélioration de l'expérience de l'usager, mais aussi dans l'amélioration du traitement de l'information en général (Zengenene, Casarosa et Meghini, 2014). La vulgarisation des processus ainsi que la capacité à communiquer les bénéfices pour l'institution sont primordiales à l'avancement afin de permettre aux instances de s'approprier les technologies du Web de données et de voir les bénéfices à court, moyen et long termes. Il peut être très difficile, voire impossible, d'évaluer les retombées économiques d'un tel projet. Il est donc nécessaire de le présenter en misant sur les autres avantages tels que l'augmentation de l'achalandage sur les plateformes Web de l'institution, la présence et la visibilité dans les résultats des moteurs de recherche, l'amélioration des services aux usagers, les possibilités de réutilisation des données permettant la création de nouvelles applications et de services et la participation à un projet d'envergure internationale et d'évolution du catalogue de bibliothèque. Les avantages présentés à la section 3.2 peuvent aussi aider à la compréhension des enjeux pour le domaine des bibliothèques et il est pertinent de les mettre de l'avant. Ces éléments doivent être présentés dans un document dans lequel on évitera le jargon technique.

6.1.2 Obtenir l'autorisation des parties prenantes

Dans une note du W3C (2014) ayant comme objectif de faciliter le développement des données gouvernementales vers les données ouvertes et liées, il s'agit de la première étape proposée. L'étape présentée précédemment est nécessaire à celle-ci, dans la mesure où elle permet de préparer les différentes instances et de présenter le projet de façon claire et simple. Une fois le document rédigé, il devient plus facile de communiquer les objectifs et d'exposer les ressources nécessaires. Pour obtenir l'autorisation des parties prenantes, il est aussi possible de présenter des exemples de projets de publication de jeux de données dans le Web de données

qui permettront de constater concrètement ce que ces applications peuvent apporter comme nouveaux services et nouvelles applications. Afin de permettre une association aux approches traditionnelles de gestion de l'information, il est possible de présenter des modèles présentant le cycle de vie des données liées et ouvertes et se baser sur la liste des étapes présentées pour permettre une vue d'ensemble.

6.1.3 Établir une licence d'utilisation

Il est primordial de statuer sur les droits d'utilisation des données publiées en indiquant qui est propriétaire des données et dans quelle mesure il est possible de les réutiliser, surtout lorsqu'il s'agit de données provenant d'institutions culturelles publiques (Bermès, 2013; Bizer et Heath, 2011; Hyvönen, 2012; Villazón-Terrazas, Vilches-Blázquez, Corcho et Gómez-Pérez, 2011; W3C, 2014-b; Zengenene, Casarosa et Meghini, 2014). Les usagers et développeurs auront d'ailleurs plus tendance à s'approprier les données si une licence présente les spécificités légales (W3C, 2014-b). Tel que présenté à la section 2.4, le mouvement des données ouvertes prend une importance de plus en plus grande et les institutions de la mémoire collective s'inscrivent graduellement dans ce courant. Mentionnons cependant qu'il est possible de lier des données sans en permettre la réutilisation – on parle alors de données liées fermées. Cette pratique est moins commune et a souvent lieu au sein d'entreprises privées. Le partage de données se fait alors, en général, uniquement à l'interne.

De plus en plus, on constate l'utilisation du modèle de licences *Creative Commons*[31]. Par exemple, Données Québec [32], Europeana, la Bibliothèque nationale d'Espagne, la Bibliothèque nationale d'Allemagne et plusieurs autres utilisent les licences *Creative Commons* pour encadrer légalement la réutilisation de leurs données. L'*Open Database License* (ODbL) est un autre exemple de licence libre qui favorise la circulation des données. Notons cependant que l'utilisation de ces licences peut être considérée comme le résultat d'un manque d'alternatives qui seraient plus adéquates pour le domaine des sciences de l'information

[31] https://creativecommons.org

[32] https://www.donneesquebec.ca/fr/licence/

(Zengenene, Casarosa et Meghini, 2014). D'autres institutions, pour leur part, rédigent simplement elles-mêmes leur licence d'utilisation ou se rattachent à une licence d'état, c'est-à-dire une licence s'appliquant à toutes les administrations gouvernementales. Dans le cadre d'un projet prenant place au sein d'une bibliothèque nationale, elle-même institution gouvernementale, cette dernière option peut s'avérer intéressante, car elle permet de suivre la ligne directrice mise en place par l'état.

Lorsque vient le temps de choisir une licence, il faut tenir compte de trois concepts différents, soit l'attribution, la restriction de l'usage commercial et la redistribution à l'identique (Bermès, Isaac et Poupeau, 2013). Les clauses d'une licence liées à l'attribution (*by*) sont celles qui précisent comment doit être reconnu le détenteur de la propriété intellectuelle de la ressource. Ainsi, dans certains cas, l'usager peut être dans l'obligation de citer la source des données dans le contexte d'une réutilisation. La restriction de l'usage commercial (*nc*) implique que le réutilisateur ne peut utiliser des données à des fins commerciales. Une licence peut aussi préciser les différentes possibilités pour l'usage commercial. Finalement, la redistribution à l'identique (*share-alike*) précise si l'usager peut ou non réutiliser et transmettre les données en utilisant une licence semblable à celle utilisée par l'institution.

6.1.3.1 Le cas de data.bnf.fr

Afin d'assurer une complète interopérabilité des données et de répondre aux objectifs du Web de données ouvert, la BnF a ouvert ses métadonnées descriptives, c'est-à-dire les données bibliographiques (plus de 12 millions de notices) et les données d'autorité (plus de 2,5 millions de notices) (BnF, 2015-h). Pour ce faire, la BnF rend ses données brutes disponibles sous la Licence ouverte française, créée par Etalab, un service du Premier ministre français faisant partie du Secrétariat général pour la Modernisation de l'Action Publique dont l'objectif est l'ouverture des données publiques et la mise sur pied d'outils veillant à assurer un gouvernement ouvert (Etalab, s.d.). La Licence ouverte implique que la « réutilisation et la reproduction des données RDF est libre et gratuite pour tout usage, y compris commercial. Une mention d'attribution est nécessaire. » (BnF, 2015-g).

6.1.3.2 Possibilités pour BAnQ

Du côté de BAnQ, il serait possible de rédiger une licence d'utilisation uniquement dans le cadre de leur projet. Cette décision implique évidemment toutes les parties prenantes et nécessite en général l'implication d'un membre de l'équipe juridique. Par exemple, comme mentionné à la section 5.2.1, dans le cadre du portail du RFN, la licence d'utilisation a été rédigée spécifiquement pour ce projet et fut inspirée de la licence du portail de données ouvertes du Québec de l'époque (RFN, s.d.-b). L'autre option qui se présenterait serait celle de simplement choisir la même licence qui est utilisée du côté de Données Québec, c'est-à-dire la licence *Creative Commons 4.0*[33]. Cette licence indique que l'usager est « (…) autorisé à :

- Partager les données ;
    - Copier, distribuer et communiquer le matériel par tous moyens et sous tous formats.
- Adapter.
    - Remixer, transformer et créer à partir du matériel pour toute utilisation, y compris commerciale. » (Creative Commons, s.d.)

Cette licence demande une attribution, c'est-à-dire que l'usager doit indiquer les crédits ainsi que si des modifications ont été effectuées ou non. Évidemment, dans la mesure où BAnQ choisirait d'utiliser cette même licence, cela implique que les ressources pour lesquelles l'institution ne possède pas les droits de diffusion ne peuvent faire partie des jeux de données publiés et ouverts.

6.1.4 Évaluer les compétences

À cette étape, il est nécessaire de commencer à évaluer quel processus de conversion sera le plus pertinent pour les jeux de données à publier (Zengenene, Casarosa et Meghini, 2014). Pour

[33] https://creativecommons.org/licenses/by/4.0/deed.fr

ce faire, une analyse de la situation qui présente les compétences des différents acteurs impliqués dans le processus doit être faite. Selon Zengenene, Casarosa et Meghini (2014), les compétences nécessaires pour les bibliothécaires et autres professionnels de l'information sont triples :

- Systèmes d'information :
    - Avoir des connaissances en ce qui a trait au téléchargement, à l'installation et à la configuration de systèmes d'information, de bases de données (particulièrement de *triplestores*) et de serveurs ;
    - Être aptes à écrire et lire les formats XML et RDF.
- Métadonnées :
    - Connaître le processus de catalogage, l'importance et la signification des métadonnées.
- Modélisation :
    - Comprendre la structure des données et être aptes à évaluer la meilleure façon de convertir les données d'une structure donnée vers RDF.

Ensuite, dans la mesure où l'institution se voit dans l'obligation d'engager des techniciens de l'externe (développeurs ou programmeurs, par exemple), les bibliothécaires et professionnels de l'information doivent être capables de communiquer leurs besoins et de décrire les ontologies et vocabulaires qu'ils souhaitent utiliser. Mettre sur pied une équipe dont l'objectif sera de mettre en place un projet de Web de données peut s'avérer particulièrement ardu. On constate d'un côté le besoin de développeurs et de programmeurs qui seront aptes à construire l'architecture, mais aussi la nécessité d'avoir au sein de l'équipe des individus qui maîtrisent parfaitement bien le contenu qui doit être manipulé. Ainsi, dans le cas qui nous intéresse, c'est-à-dire dans le cadre d'une institution documentaire, la place de professionnels de l'information au sein de l'équipe de développement est absolument nécessaire. Les compétences doivent donc être hétérogènes, diversifiées et bien développées.

6.1.4.1 Le cas de data.bnf.fr

Du côté de la BnF, il est difficile de se prononcer sur ce sujet, dans la mesure où nous n'avons pu obtenir d'informations exactes en ce qui a trait aux membres de l'équipe de développement du projet data.bnf.fr. Comme mentionné à la section 5.1.1.3, une équipe de six personnes y travaillent à temps plein depuis le début du projet, mais nous n'avons pu avoir accès aux spécificités et aux compétences de celles-ci. Étant donné que la majorité du travail de développement s'effectue du côté du prestataire Logilab, on peut déduire qu'il n'a pas été nécessaire d'engager ou d'affecter des employés de la BnF au travail de codage. Cependant, la BnF nous indiquait qu'il avait été relativement difficile pour les membres du Département des Systèmes Informatiques d'y travailler parce qu'ils étaient habitués de travailler en langage Java alors que les applications développées par Logilab sont en langage Python. Elle indiquait ainsi qu'il aurait été plus facile de développer le projet en collaboration avec le prestataire s'il y avait eu davantage de cohérence dans les technologies utilisées par les deux parties. Il est donc préférable de s'assurer qu'il y ait une entente à ce niveau ou encore que tous les participants au projet maîtrisent du moins minimalement les technologies utilisées par chacun.

6.1.4.2 Possibilités pour BAnQ

Du côté de BAnQ, il est encouragé que, dans la mesure du possible, le développement d'un projet de Web de données soit fait à l'interne dans son entièreté et que l'on ne fasse pas affaire avec un prestataire externe afin d'éviter des coûts trop importants. Mentionnons la présence d'un seul et unique bibliothécaire de métadonnées au sein de l'équipe qui devrait inévitablement être impliqué dans le travail de développement. D'abord parce qu'il est spécialisé dans la gestion de métadonnées et ensuite parce qu'il a une formation en sciences de l'information, ce qui lui permet de bien maîtriser le contenu. Au sein de l'équipe de la Direction générale des technologies de l'information et des télécommunications, on retrouve aussi une analyste informatique aussi spécialiste en préservation et diffusion de collections numériques qui devra aussi jouer un rôle majeur dans le développement. Il faut aussi tenir compte de la participation d'un architecte de données. Mise à part la constitution de l'équipe de

développement, il est nécessaire de mentionner le fait qu'un tel projet touchera de près ou de loin d'autres départements, entre autres la Direction de la numérisation, la Direction du traitement documentaire des collections patrimoniales et la Direction de la Collection nationale et des collections patrimoniales.

6.1.5 Évaluer les jeux de données

En ce qui a trait à l'identification des jeux de données à publier, on doit analyser lesquels on souhaite rendre disponibles, les formats dans lesquels ils sont encodés ainsi que les technologies nécessaires à chaque situation. On recommande en général de prioriser la publication de données qui sont uniques à l'institution et qui présentent un intérêt de réutilisation par des usagers. Trois types de données sont à considérer (Bermès, Isaac et Poupeau, 2013) :

- Notices se trouvant dans le catalogue ou la base de données ;
- Référentiels ou notices d'autorité ;
- Éléments de métadonnées ou ontologies développées.

Les référentiels ou les données d'autorité présentent pour leur part un avantage intéressant, dans la mesure où ils sont souvent le résultat d'efforts importants de normalisation et offrent une grande possibilité de réutilisation par d'autres institutions. Par exemple, comme mentionné précédemment, les thésaurus et listes d'autorité RAMEAU (BnF) et Library of Congress Subject Headings (LCSH) ont été rapidement publiés sur le Web de données. Pour ce qui est de la troisième option, soit les ontologies ou les éléments de métadonnées développés au sein de l'institution, il est plutôt recommandé d'utiliser des ontologies existantes plutôt que de procéder à la création de celles-ci afin d'assurer une meilleure interopérabilité. Il peut cependant arriver que certaines institutions ressentent le besoin de créer leur ontologie lorsque leurs données présentent des spécificités particulières. Finalement, il est intéressant d'envisager la publication de données administratives et techniques afin de s'inscrire dans le mouvement des données ouvertes, telles que des données de nature statistique et transactionnelle. Dans tous les cas, il est important de statuer quant au choix des jeux de données à publier, car cela facilitera les tâches suivantes et l'analyse du travail qui sera à effectuer.

6.1.5.1 Le cas de data.bnf.fr

Pour sa part, data.bnf.fr regroupe les ressources du catalogue général, de la base de données d'archives et de manuscrits et de Gallica. On retrouve aussi la publication de données provenant de partenaires institutionnels et de sites externes. Le processus de publication est toujours en cours et, en novembre 2014, on retrouve 60 % des catalogues, c'est-à-dire 630 000 auteurs, 173 000 thèmes et plus de 7 millions de documents (BnF, 2015-g). Comme indiqué ci-dessus, les données RAMEAU sont aussi disponibles, ainsi que l'ontologie BnF, développée pour répondre à des besoins spécifiques. Ces besoins sont la nécessité d'exprimer (BnF, 2016) :

- Le numéro ISBN ;
- Un texte alternatif pour les images ;
- Le numéro d'identification *European Article Numberting* (EAN) ;
- La cote d'un document d'archives ;
- La vignette préférée d'une image ;
- Le numéro *International Standard Serial Number* (ISSN) ;
- Le numéro *International Standard Music Number* (ISMN) ;
- L'URL vers une exposition virtuelle de la BnF ;
- La notice analytique ;
- L'édition d'un ouvrage destinée à un public jeune ;
- Le code de fonction de rôle de la personne ou de l'organisation en relation avec l'ouvrage écrit ;
- Le nom pour désigner le rôle des contributeurs.

La BnF respecte les recommandations à ce niveau, car elle travaille à rendre le plus grand nombre de jeux de données disponibles et a mis de l'avant dès le départ un jeu de données spécifique à son institution, soit RAMEAU.

6.1.5.2 Possibilités pour BAnQ

Pour sa part, BAnQ, en sa qualité de bibliothèque nationale, a en sa possession une grande quantité de ressources patrimoniales au sein de son catalogue qui constituent une excellente source de jeux de données uniques et qui ont un intérêt de réutilisation. Ensuite, les notices de la Bibliographie du Québec (BQ) sont, pour leur part, déjà accessibles via le site Internet du gouvernement ouvert du Québec. Ces données sont propres à BAnQ et n'impliquent aucune problématique en ce qui a trait aux droits de diffusion. Pour d'autres institutions documentaires, ces données peuvent avoir un bon potentiel de réutilisation et peuvent être une source de création d'usages innovants du côté des usagers. Étant donné que BAnQ n'est pas détenteur des droits du Répertoire de vedettes-matière (RVM) qu'elle utilise (le RVM étant un service payant offert par l'Université Laval), elle n'est pas en mesure de rendre disponibles ces informations en données ouvertes et liées. Cependant, BAnQ est l'auteure de ses propres notices de fichiers d'autorité, donc il serait possible d'envisager de les rendre disponibles. Comme indiqué à la section 4.3.2.3, l'institution déverse d'ailleurs déjà ces notices vers VIAF.

Finalement, BAnQ propose aussi différents services qui lui sont propres tels que le Bottin des éditeurs canadiens-français, l'Enquête annuelle des bibliothèques publiques du Québec ou le Réseau de diffusion des archives du Québec, qui impliquent aussi de nombreuses données pertinentes pour d'autres acteurs. Il serait aussi possible de rendre certaines données administratives disponibles telles que les statistiques de fréquentation et de prêts. Les possibilités sont donc nombreuses pour BAnQ et plusieurs options s'offrent à elle quant au choix des jeux de données à publier.

6.1.6 Choisir le modèle de publication et évaluer les outils nécessaires

La figure 20 (Bizer et Heath, 2011) permet une visualisation des modèles de publication les plus répandus. Cette étape est importante dans la mesure où elle permet d'avoir une idée d'ensemble du processus de publication des données. D'abord, si les documents sont en format texte en langue naturelle (p. ex., rapports ou articles), il est possible d'utiliser un extracteur

d'entités nommées et d'annotations Web sémantique, qui permettra la reconnaissance et l'extraction d'informations dans un corpus donné et l'annotation en RDF (Bizer et Heath, 2011). Le fonctionnement de ce type d'extracteur est le suivant : les documents sont annotés par l'extracteur à l'aide des URI des entités nommées dans le document. La publication de ces documents accompagnés de ces annotations augmente les chances de découverte et permet aux applications du Web sémantique d'effectuer des liens vers ceux-ci. Notons cependant que les extracteurs d'entités nommées ne permettront pas d'extraire les relations entre les entités. Ainsi, on aura la possibilité de définir les URI, mais non les triplets RDF. Un travail supplémentaire est donc nécessaire afin d'obtenir des fichiers RDF qui pourront être versés sur un serveur Web tel que Apache. Cette façon de faire est la plus simple et est principalement utilisée lorsque la quantité de fichiers est limitée, car le travail se fait principalement de façon manuelle.

Ensuite, dans la mesure où les données sont déjà structurées, ce qui est le cas dans la plupart des situations liées au domaine des sciences de l'information, tout dépend alors de la façon dont les données sont stockées. Ainsi, si les données sont stockées dans une base de données relationnelles, deux technologies de conversion peuvent être utilisées : RDB-to-RDF[34] ou l'utilisation d'un système de gestion de contenu qui permet d'exprimer les données en RDFa (voir section 2.6). Un système de gestion de contenu permet la création et la gestion de contenu numérique via une interface conviviale. Ainsi, certains outils comme Drupal[35], lui-même un système de gestion de contenu, permettent donc de transformer les données structurées qui se trouvent dans sa base de données en RDFa en ajoutant des attributs RDF aux éléments HTML. Mentionnons aussi l'existence des langages de *mapping* qui permettent de transformer les données provenant de bases de données relationnelles en RDF[36]. Lorsque les données sont encodées dans des bases de données relationnelles, il est en général préférable de ne pas procéder à un transfert des données vers un *triplestore* afin de conserver l'infrastructure de gestion de l'information déjà mise en place (Bizer et Heath, 2011).

[34] *Relational database to RDF*

[35] https://www.drupal.org/

[36] D2RQ ou R2RML, par exemple

Il est aussi possible d'accéder aux données stockées via une interface de programmation applicative (API). Dans ce cas-ci, la situation est plus complexe, dans la mesure où l'institution devra développer un adaptateur personnalisé (*wrapper*). Un adaptateur permettra de régler certaines limites qui sont associées aux API, comme, par exemple, le fait que leur contenu ne peut être repéré par les moteurs de recherche. En général, les adaptateurs permettent d'assigner des URI aux ressources sur lesquelles l'API fournit des données, de reformuler une requête pour que celle-ci soit comprise par l'API et de transformer les résultats de cette requête en RDF (Bizer, Cyganiak et Heath, 2007; Bizer et Heath, 2011). Le développement d'un tel outil implique des connaissances informatiques qui dépassent en général celles déjà acquises par les professionnels de l'information, donc l'appel à une ressource externe peut alors être nécessaire.

Finalement, les données structurées peuvent aussi être stockées dans des *triplestores*. La plupart des *triplestores* offrent une interface de données liées, ce qui peut faciliter grandement le travail de programmation. On peut donc accéder directement à un point d'accès SPARQL et il est possible d'effectuer la configuration et la gestion du contenu à même le *triplestore*.

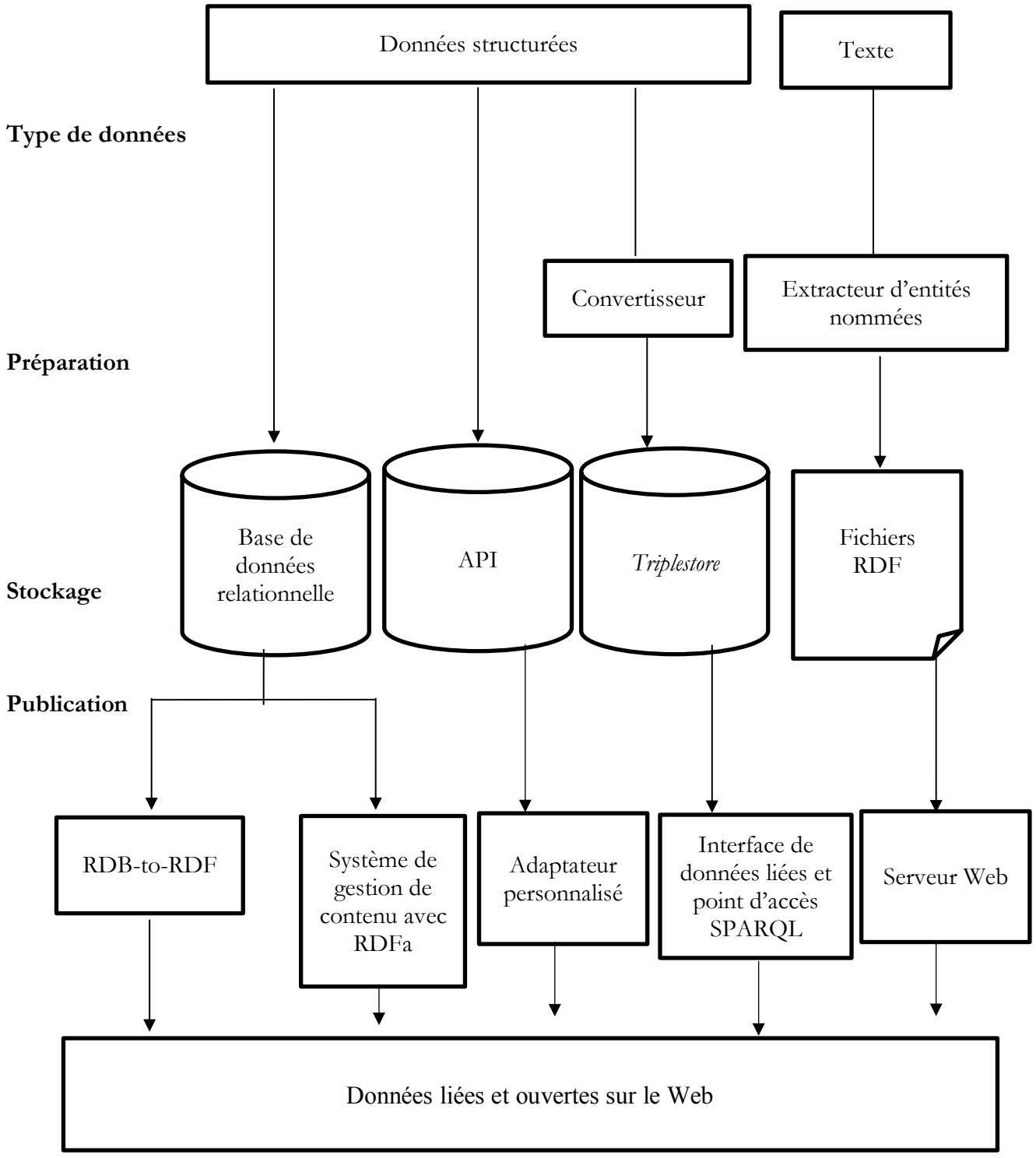


Figure 20. Modèles les plus répandus de publication de données liées et ouvertes sur le Web (Bizer et Heath, 2011, p.70, notre traduction)

6.1.6.1 Le cas de data.bnf.fr

Comme présenté à la section 5.1.1.4, la BnF a choisi de travailler à partir d'une base de données relationnelles plutôt qu'à partir d'un *triplestore*. Cette décision s'explique entre autres par le fait que les travaux autour de data.bnf.fr ont débuté en 2011. À ce moment, les technologies relatives au Web de données étaient moins développées, la documentation sur le sujet était plus rare et les risques étaient relativement élevés. Ainsi, afin d'éviter des coûts trop élevés et des difficultés quant à la compréhension de cette nouvelle façon de faire par les membres de l'équipe, la décision a été prise de se concentrer sur une technologie déjà maîtrisée. Cet exemple démontre bien que le choix d'utiliser un *triplestore* n'est pas obligatoire et qu'il est possible de travailler à partir d'une base de données relationnelles. L'option choisie par la BnF est donc tout à fait fonctionnelle et se caractérise aussi par le fait que le développement s'est fait du côté de son prestataire Logilab.

6.1.6.2 Possibilités pour BAnQ

Les deux options les plus plausibles pour BAnQ seraient celles de convertir les données à partir d'une base de données relationnelles ou d'aller vers l'option du *triplestore*. Étant donné la grande quantité d'informations, le choix doit se faire principalement en fonction de la performance à long terme ainsi que de la facilité d'utilisation. Le choix du *triplestore* doit être l'objet de discussions, car le nombre d'outils disponibles est important et chacun a ses avantages et désavantages. En général, il est préférable d'aller vers un logiciel libre et gratuit plutôt qu'un logiciel propriétaire afin d'éviter des coûts trop importants au moment de l'achat ainsi que dans le futur et la désuétude possible de la technologie. De plus, le fait de choisir l'option du logiciel libre s'inscrit dans la Politique-cadre sur la gouvernance et la gestion des ressources informationnelles des organismes publics du Gouvernement du Québec[37] ainsi que dans la

[37] http://www.tresor.gouv.qc.ca/fileadmin/PDF/ressources_informationnelles/Politique__lois/politique_cadre.pdf

Stratégie gouvernementale en technologies de l'information[38]. Parmi les plus utilisés, on retrouve Jena[39], Sesame[40] et OpenLink Virtuoso[41]. Notons tout de même que certains systèmes présentent une meilleure performance, tel que Oracle[42].

6.1.7 Attribuer les URI

Comme mentionné précédemment (section 2.5.1), il est nécessaire de porter une attention particulière à la création d'URI pour décrire les ressources et les relations qui les unissent. Cette étape permet aussi de constater la quantité d'entités avec lesquelles on devra travailler et qu'il sera nécessaire d'identifier. Les URI se doivent d'être construits de façon simple, stable, pérenne et gérable (Berners-Lee, 2008; Villazón-Terrazas, Vilches-Blázquez, Corcho et Gómez-Pérez, 2011).

On devra d'abord choisir un nom de domaine qui ne sera pas susceptible de changer au fil du temps. Ensuite, pour ce qui est de la description des ressources, il est préférable d'utiliser des identifiants déjà existants, tels que les référentiels ou listes d'autorité, et qui ont démontré leur persistance dans le temps. Par exemple, un ISBN pour un livre, des codes de bibliothèque ou, comme c'est le cas pour la BnF comme nous le verrons ci-dessous, le numéro de notice associée à la ressource (Bermès, Isaac et Poupeau, 2013; Villazón-Terrazas, Vilches-Blázquez, Corcho et Gómez-Pérez, 2011; Zengenene, Casarosa et Meghini, 2014). Dans la mesure où aucun identifiant n'existe, l'institution devra en créer de nouveaux et deux options s'offrent alors à elle. L'utilisation d'identifiants opaques, qui n'ont aucune signification particulière, ou d'identifiants signifiants, qui présentent une information lisible et compréhensible par l'humain (voir section 2.5.1). Selon certains auteurs, il est en général préférable d'utiliser des identifiants

[38] http://www.tresor.gouv.qc.ca/fileadmin/PDF/ressources_informationnelles/strategie_ti/strategie_ti.pdf

[39] http://jena.apache.org/

[40] http://rdf4j.org/

[41] http://virtuoso.openlinksw.com/

[42] http://www.oracle.com/technetwork/database/options/spatialandgraph/overview/rdfsemantic-graph-1902016.html

signifiants (Stuart, 2011; Villazón-Terrazas, Vilches-Blázquez, Corcho et Gómez-Pérez, 2011), car ils permettent de faciliter les tâches relatives à la maintenance et à l'exploitation des URI. Notons cependant que les identifiants opaques présentent aussi certains avantages, soit l'unicité, la pérennité, le fait qu'ils ne seront ainsi pas affectés par les changements d'environnements informatiques et, finalement, qu'ils permettent l'identification de la ressource indépendamment de la langue utilisée. Bermès, Isaac et Poupeau (2013) indiquent d'ailleurs qu'il est préférable d'utiliser des identifiants opaques lorsqu'on travaille avec des ensembles importants et hétérogènes de ressources. Comme indiqué à la section 2.5.1, il est important de s'assurer de respecter le concept de négociation de contenu et de procéder à la création d'au moins trois URI pour identifier la même ressource (ou ses différentes représentations).

Ainsi, il est nécessaire d'accorder une attention particulière à l'origine de l'identifiant (référentiel ou liste d'autorité), à la forme de l'identifiant (opaque ou signifiant) et à la création d'au moins trois identifiants pour une même ressource (un pour le concept, un pour la représentation HTML et un pour la représentation RDF).

6.1.7.1 Le cas de data.bnf.fr

Pour sa part, data.bnf.fr propose par exemple les URI suivants (l'exemple est inspiré de Bermès, Isaac et Poupeau, 2013) :

- Adresse de la page HTML :
    - http://data.bnf.fr/11900875/rejean_ducharme/
- URI du concept :
    - http://data.bnf.fr/ark:/12148/cb11900875z
- URI de la personne :
    - http://data.bnf.fr/ark:/12148/cb11900875z#foaf:Person
- Adresse du flux RDF (XML) pour l'interprétation par la machine :
    - http://data.bnf.fr/11900875/rejean_ducharme/rdf.xml
- Adresse de la page du catalogue :
    - http://catalogue.bnf.fr/ark:/12148/cb11900875z

On constate donc que la partie qui identifie Réjean Ducharme et qui présente une unicité et une pérennité est une suite de caractères qui est celle du numéro de la notice que l'on peut repérer dans la page du catalogue, numéro lui-même déjà existant (11900875). La BnF a donc fait le choix d'utiliser des identifiants opaques plutôt que signifiants, ce qui permet une unicité et une pérennité. De plus, la BnF crée ses URI grâce à des identifiants pérennes (identifiants ARK). Aussi, comme nous l'avons vu à la section 2.5.1, il y a respect des pratiques exemplaires par la création d'au moins trois URI, soit un pour la page HTML, un pour le concept et un pour l'adresse du flux RDF. L'institution respecte aussi le concept de négociation de contenu et utilise un mécanisme de redirection (HTTP 303 présenté à la section 2.5.1) pour rediriger le demandeur (humain ou machine) vers la bonne page.

6.1.7.2 Possibilités pour BAnQ

BAnQ stocke ses ressources numériques dans un entrepôt DSpace[43]. DSpace est un logiciel libre dont l'objectif est de permettre la publication et le stockage de contenu numérique. BAnQ l'utilise comme entrepôt et comme passerelle vers son site Web pour la diffusion de son contenu numérique patrimonial. L'un des avantages du logiciel DSpace est le fait qu'il génère, pour chaque ressource, un identifiant ARK, donc opaque et pérenne. Cela peut donc constituer un bon point de départ pour la création d'URI, non pas seulement pour les ressources numériques en tant que telles, mais aussi pour les personnes, les lieux et les concepts qui sont en lien avec celles-ci.

6.1.8 *Choisir son modèle de données et faire le* mapping

Cette étape, qui consiste à développer la structure sémantique, peut présenter certaines difficultés, surtout dans la mesure où les métadonnées sont encodées de manières différentes, dans le cas où, par exemple, elles proviendraient de catalogues différents. C'est le cas pour BAnQ et la BnF, qui possèdent des notices encodées en MARC et en EAD. Un modèle de

[43] http://www.dspace.org/

données découle lui-même de la modélisation des données, c'est-à-dire le fait d'analyser et de concevoir la structure des données qui sont contenues dans un système, mais aussi la sémantique des données (Bermès, Isaac et Poupeau, 2013). D'ailleurs, la modélisation des données dépendra du mode de publication choisi (section 6.1.13) et de la nature des données. Ainsi, cette étape implique les choix relatifs aux classes et aux propriétés qui seront sélectionnées pour décrire les entités et les relations qui les unissent. Pour ce faire, il est donc important de bien connaître le niveau de description et les éléments de métadonnées utilisés pour, par la suite, effectuer un *mapping* entre les jeux de données. La plupart du temps, l'institution devra utiliser plusieurs ontologies et vocabulaires pour représenter son information (voir section 2.6). Bermès, Isaac et Poupeau (2013, p. 91) indiquent :

> La construction du modèle de données correspond donc au fait de choisir dans les ontologies existantes les classes et les propriétés qu'on souhaite utiliser, vérifier leurs relations (hiérarchies de classes et sous-classes, de propriétés et sous-propriétés, domaines et codomaines, liens avec d'autres ontologies), et enfin compléter ce modèle en créant, si nécessaire, les éléments de métadonnées manquants ou les relations avec des ontologies existantes.

La difficulté réside aussi dans le fait que le modèle de données doit permettre de représenter toutes les entités et leurs relations, tout en assurant une logique durant le processus. Le choix des ontologies et des vocabulaires est donc l'une des parties les plus importantes du processus de publication de données liées et ouvertes sur le Web et on encourage la réutilisation de standards (Villazón-Terrazas, Vilches-Blázquez, Corcho et Gómez-Pérez, 2011; W3C, 2014-b; Zengenene, Casarosa et Meghini, 2014). Hyvönen (2012) présente dans son ouvrage *Publishing and Using Cultural Heritage Linked Data on the Semantic Web* une liste des schémas de métadonnées, des ontologies et des vocabulaires disponibles. Cette liste ne sera pas présentée dans son intégralité ici, mais il s'agit d'une référence intéressante lorsque vient le moment de procéder à la sélection. Le tableau 8 présente toutefois certains schémas de métadonnées, ontologies ou vocabulaires qui peuvent répondre à certains besoins des institutions documentaires.

Tableau 8. Exemples de schémas de métadonnées, ontologies et vocabulaires

Schéma de métadonnées, ontologie ou vocabulaire	Description
Visual Resource Association (VRA) *Core*	Éléments pour la description d'images d'art et d'architecture
Categories for the description of works of art (CDWA)	Cadre d'interopérabilité pour la description d'images d'art et d'architecture
Europeana Data Model (EDM)	Cadre d'interopérabilité pour la description de ressources patrimoniales numérisées
CIDOC Conceptual Reference Model (CIDOC CRM)	Modèle de référence pour la description du patrimoine culturel
Light Weight Information Describing Objects (LIDO)	Schéma pour la description d'objets muséaux

Nous présenterons plutôt les considérations dont il faudra alors tenir compte. Notons que ce choix dépend grandement des objectifs de l'institution ainsi que le type de données qu'elle souhaite publier. De plus, il est fort probable que le *mapping* ait comme conséquence la perte de granularité au niveau des descriptions. Cet aspect peut décevoir, mais il est à noter que les données qui seront publiées dans le Web de données n'ont pas comme objectif de remplacer les notices bibliographiques qui sont souvent plus précises, mais plutôt d'apporter une couche supplémentaire d'interopérabilité et une possibilité de réutilisation. Le W3C (2014) une liste visant à permettre une évaluation des différents vocabulaires et ontologies envisagés. Les points de cette liste sont les suivants :

- Les vocabulaires et ontologies doivent être documentés;
    - C'est-à-dire qu'il doit être possible d'accéder à des pages lisibles par l'humain qui présentent le vocabulaire ou l'ontologie ainsi que les classes et propriétés qui la constituent.
- Les vocabulaires et ontologies doivent être autodescriptifs;
    - C'est-à-dire que chaque propriété ou terme du vocabulaire ou de l'ontologie doit présenter de l'information de façon explicite dans sa représentation. Ainsi, on retrouve un *label*, une définition et un commentaire présentant ou des exemple(s) pour chaque propriété ou terme.
- Les vocabulaires et ontologies doivent être décrits dans plus d'une langue;

    - C'est-à-dire que l'utilisation doit être possible dans le cadre d'une situation multilingue.
- Les vocabulaires et ontologies doivent être utilisés par d'autres jeux de données publiés;
    - Il s'agit là d'un aspect important dans la mesure où les avantages du Web de données résident principalement dans ce concept de réutilisation.
- Les vocabulaires et ontologies doivent être pérennes;
    - Cet aspect implique une relation de confiance et les vocabulaires et ontologies doivent assurer une certaine garantie de durée et de maintenance dans le temps.
- Les vocabulaires et ontologies doivent être publiés par des organisations de confiance;
    - Il est essentiel de vérifier la source.
- Les vocabulaires et ontologies doivent être représentés par des URI pérennes;
    - On permet ainsi une réutilisation facile et garantie dans le temps.
- Les vocabulaires et ontologies doivent fournir une politique présentant les versions;
    - Des changements majeurs doivent être documentés et cette documentation doit être accessible.

Cette liste permet un certain contrôle lors de la réutilisation de vocabulaires et d'ontologies. Dans la mesure où l'institution souhaite créer son propre outil, ces points sont aussi à considérer. Finalement, en ce qui a trait à l'utilisation de SKOS pour la publication d'une liste de termes ou d'une taxonomie ou lorsque OWL n'est pas adéquat pour une situation en particulier (W3C, 2014-b), de nouvelles règles s'appliquent. Celles-ci ne seront pas abordées ici étant donné leur complexité, mais il est intéressant de savoir que cette possibilité est présente et que la documentation existe pour aider les institutions dans ce processus.

Ensuite, une fois le(s) vocabulaire(s) et ontologie(s) choisis et ainsi, le modèle de données défini, il est nécessaire de procéder à la conversion des données grâce au processus de *mapping*. Il s'agit ici d'une sous-étape qui peut donner du fil à retordre dans la mesure où elle nécessite une bonne maîtrise des technologies qui seront utilisées, du vocabulaire ou de l'ontologie et des formats. En général, on commence par la création un tableau de conversion qui permettra d'indiquer la source, la cible, la règle de *mapping* et, si possible, un exemple. Il est possible de

créer la conversion selon différents niveaux de précision et de détail, le choix revenant à l'institution d'évaluer les différentes possibilités. Puis, on doit par la suite transformer les données en fonction des règles de *mapping*. Pour ce faire, on utilise généralement un programme informatique et le choix de cet outil dépend du format source, comme présenté à la section 6.1.6.

6.1.8.1 Le cas de data.bnf.fr

La figure 21 présente une version simplifiée du modèle de données de data.bnf.fr (BnF, 2016).

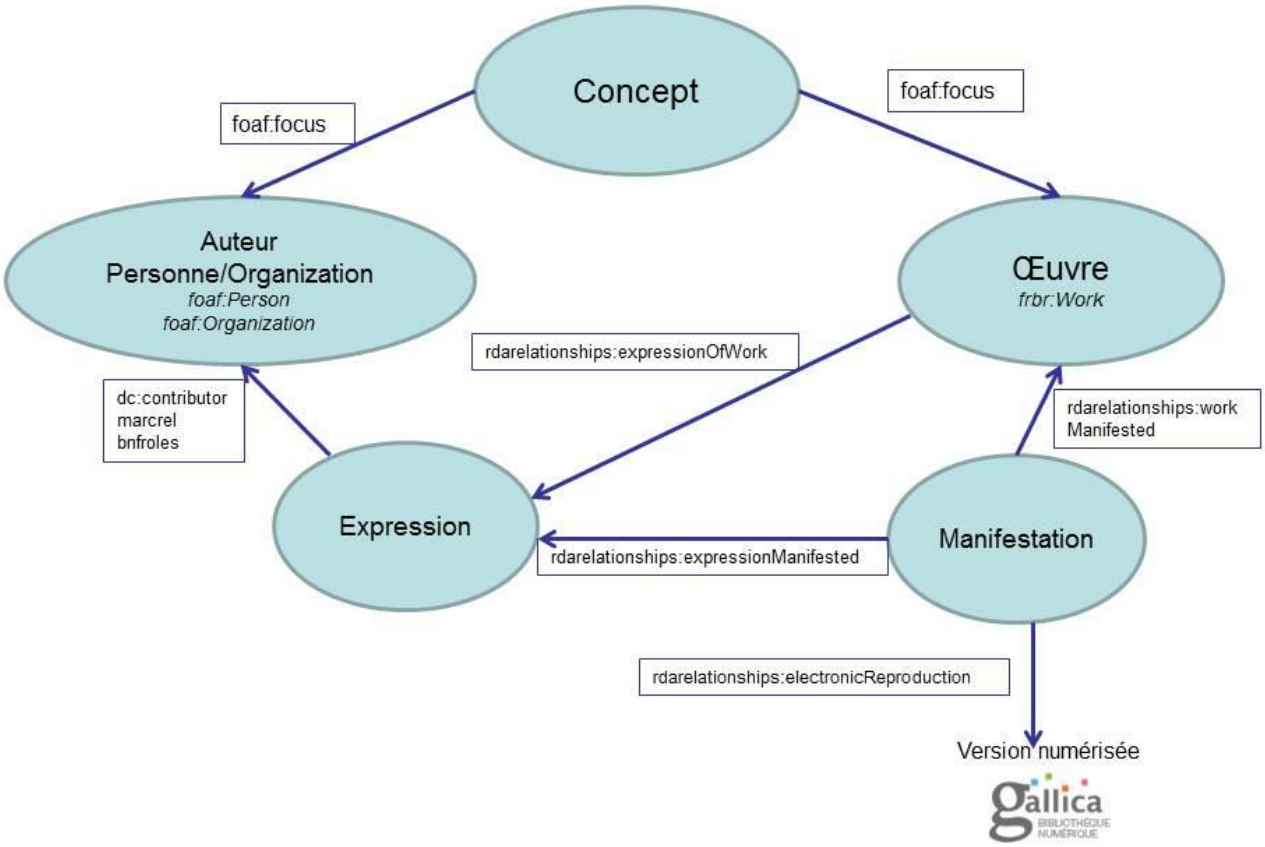


Figure 21. Modèle de données simplifié de data.bnf.fr

Comme on peut le constater, le modèle est basé sur le modèle FRBR présenté précédemment. On y retrouve aussi l'utilisation du schéma de métadonnées Dublin Core, du vocabulaire RDA et du vocabulaire FOAF. Cependant, data.bnf.fr utilise aussi des référentiels spécifiques de la BnF tels que les codes de pays, les codes de rôles, les types de sujets RAMEAU, les codes de classement géographique et les genres musicaux. On retrouve aussi l'ontologie BnF. Le tableau 9 présente l'ensemble des vocabulaires et ontologies externes utilisés par data.bnf.fr (BnF, 2016).

Tableau 9. Vocabulaires, schémas de métadonnées et ontologies externes utilisées par data.bnf.fr

Vocabulaire ou ontologie	**Préfixe**
The Bibliographic Ontology	bibo
BIO : a vocabulary for biographical information	bio
DCMI Metadata Terms	dc
DCMI Box Encoding Scheme	dcmi-box
DCMI Metada Terms	dcterms
Friend of a friend	foaf
FRBR entities from RDA element set	frbr-rda
WGS84 Geo Positioning	geo
GeoNames Ontology	geonames
Ontologie des unités administratives de l'IGN	ign
Données géographiques de l'Insee	insee
International Standard Name Identifier	isni
MARC Relator Terms	marcrel
Music Ontology	mo
ORE Specification	ore
Web Ontology Language	owl
RDA Group 1 Elements	rdagroup1elements
RDA Group 2 Elements	rdagroup2elements
RDA Relationships for Works, Expressions, Manifestations, Items	rdarelationships
RDF Schema	rdfs
Simple Knowledge Organization System	skos

L'une des particularités de data.bnf.fr est le fait que les auteurs et les œuvres, considérés comme les entités principales, sont définis comme entités du monde réel, mais aussi comme concept abstrait grâce à l'utilisation de skos:Concept (Bermès, Isaac et Poupeau, 2013; BnF, 2016). Cette distinction permet donc de présenter un auteur comme créateur, mais aussi comme sujet, ce qui permet de garder une logique et d'aligner entre eux les différents concepts grâce aux propriétés de SKOS.

De plus, data.bnf.fr rend accessibles sur son site Web toutes ses tables de *mapping* (MARC et EAD), ce qui permet une réutilisation éventuelle par une autre institution et une meilleure compréhension du processus. Le tableau 10 montre un extrait du *mapping* EAD vers RDF et le tableau 11 montre un extrait du *mapping* des notices bibliographiques en INTERMARC (le format de travail utilisé par la BnF) vers RDF.

Tableau 10. Extrait du *mapping* EAD vers RDF de la BnF

Libellé du catalogue	**Élément EAD dans sa hiérarchie**	**Correspondance RDF**
Titre d'œuvre	<ead><archdesc><did><unittitle><title>	dc:title
Intitulé archivistique	<ead><archdesc><did><unittitle><	dc:title
Cote	<ead><archdesc><did><unitid type="cote">	bnf-onto:cote
Date	<ead><archdesc><did><unitdate>	dc:date
Document numérisé	<ead><archdesc><dao href="permalien de Gallica"/>	rdarelationships:electronicReproduction

Tableau 11. Extrait du *mapping* INTERMARC vers RDF de la BnF

Libellé du catalogue	**Zone INTERMARC**	**Correspondance RDF**	**Commentaire**
Titre	245 $a	dcterms:title	type=frbr:manifestation
Publication	260 $a $c $d	dcterms:publisher	type=frbr:manifestation
ISBN	020 $a	bnf-onto:ISBN	type=frbr:manifestation
Langue	008 position 31-33	rdagroup1elements:Note	type=frbr:expression

Année de publication	260 $d	rdagroup2elements: dateOfPublicationManifestation	Aligné sur la page data.bnf.fr de type « date »
A pour sujet	600, 606, 610, 616, 617	dcterms:subject	Pointe vers un skos:Concept

On remarque l'utilisation des vocabulaires, schémas de métadonnées et ontologies mentionnés précédemment et les commentaires relatifs au modèle FRBR. Cette façon de faire a démontré sa faisabilité.

6.1.8.2 Possibilités pour BAnQ

Les efforts effectués du côté de BnF en ce qui a trait au *mapping* sont intéressants pour BAnQ dans la mesure où elle a aussi en sa possession des données encodées en MARC et en EAD. Cet exemple est pertinent, car il s'agit d'une occasion pour BAnQ de s'inspirer des tableaux de conversion rendus disponibles par la BnF. Ce travail peut être ardu parce qu'il implique de prendre connaissance de toutes les possibilités pour toutes les ressources décrites. Ainsi, il faudra prévoir des une équipe ayant des connaissances approfondies du contenu qu'a BAnQ en sa possession afin de créer des tableaux de conversion correspondant, tout en s'assurant d'éviter une perte de granularité trop importante. Il est à envisager d'incorporer le modèle FRBR au sein du catalogue afin de faciliter la conceptualisation et la transposition des informations vers le Web de données.

6.1.9 Nettoyer les données

Avant de convertir les données en RDF, il est nécessaire de s'assurer qu'on y retrouve le moins d'erreurs possibles et qu'on en augmente la qualité. On pense à des erreurs typographiques, des valeurs manquantes, des doublons ou des contradictions (Rahm et Do, 2000; Van Hooland et Verborgh, 2014). Par exemple, il pourrait y avoir des informations manquantes en ce qui a trait à une notice bibliographique (date de publication, numéro de l'édition, etc.) ou la forme choisie pour le nom de l'auteur pourrait ne pas correspondre à la notice d'autorité. Il est évident que le résultat ne sera pas parfait, mais il est toujours préférable

de faire quelques vérifications avant de procéder à la conversion. Il est donc nécessaire de s'assurer que les données sont dans un format structuré et on peut par la suite utiliser des outils tels que OpenRefine[44] ou DataWrangler[45] pour procéder au nettoyage.

6.1.9.1 Le cas de data.bnf.fr

Nous n'avons malheureusement pas pu obtenir d'informations précises quant au mode de nettoyage des données privilégié par la BnF.

6.1.9.2 Possibilités pour BAnQ

Afin de faciliter le processus, le meilleur moyen serait d'utiliser un outil gratuit tels que OpenRefine mentionné précédemment qui permettrait de procéder à la vérification de la qualité des données. Si l'institution se sent confiante quant à celles-ci, il est possible de sauter cette étape, mais si on découvre des lacunes au sein des jeux de données après les étapes suivantes, le travail de nettoyage sera plus long et ardu. C'est pourquoi il est recommandé d'effectuer une vérification dans tous les cas.

6.1.10 Enrichir les données en faisant des liens

Cette étape est primordiale et a comme objectifs de définir des triplets qui seront connectés à d'autres et créer des triplets qui permettront de définir des relations, le tout à l'interne et à l'externe (Zengenene, Casarosa et Meghini, 2014). Par exemple, comme démontré à la section 2.5.1, il est possible de lier l'URI d'un auteur à d'autres notices d'autorité d'institutions reconnues comme la LC ou VIAF. Lorsqu'on crée des liens au sein même du jeu de données, il est nécessaire de s'assurer que toutes les ressources seront connectées entre elles. Pour ce qui est des liens vers l'externe, il est préférable de se lier à des jeux de données qui sont

[44] http://openrefine.org

[45] http://vis.stanford.edu/wrangler/

fiables, grandement utilisés, stables et bien établis. C'est le cas de DBpedia, Geonames, Europeana, VIAF et Library of Congress.

Le choix de ces jeux de données dépend de plusieurs facteurs, tels que la valeur ajoutée que ceux-ci peuvent apporter aux données de l'institution, mais aussi la visibilité que ces liens peuvent apporter. Notons la nécessité de prendre connaissance de la licence d'utilisation de chacune des entités externes vers lesquelles on souhaite effectuer des liens et, ainsi, réutiliser les données. Selon Bermès, Isaac et Poupeau (2013), la création de liens implique trois tâches liées, soit :

- Identifier quels sont les points de contact entre ses propres jeux de données et ceux de l'externe;
- Identifier les liens qui unissent les entités de l'externe et celles de son jeu de données;
- Évaluer quelle est la meilleure méthode pour effectuer ces liens, soit de façon manuelle ou automatique.

Pour procéder aux alignements entre listes d'autorité et vedettes-matière, SKOS (skos:exactMatch, skos:closeMatch, etc.), OWL (owl:equivalentClass, owl:sameAs, etc.) et RDFS (rdfs:seeAlso) sont souvent utilisés lorsqu'il est nécessaire de démontrer une équivalence entre un identifiant créé par l'institution et un identifiant provenant de l'externe.

6.1.10.1 Le cas de data.bnf.fr

La figure 22 présente comment les données de data.bnf.fr sont alignées à des données équivalentes externes (BnF, 2016). Elle présente les entités Lieu, Auteur, Date et Concept. Ainsi, pour le Lieu, on constate un lien vers GeoNames. Les auteurs, pour leur part, sont liés à VIAF, IdRef et DBpedia. Les dates sont aussi liées à DBpedia. Finalement, les sujets RAMEAU, considérés comme des concepts (skos:Concept) sont liés aux référentiels de la Library of Congress, de la Bibliothèque nationale allemande, au Thésaurus W (Thésaurus pour la description et l'indexation des archives locales anciennes, modernes et contemporaines), à

GeoNames, au thésaurus Agrovoc, à Dewey, à l'*International Standard Name Identifier* (ISNI) lorsqu'il s'agit d'un auteur et à l'Insee lorsqu'il s'agit d'un lieu.

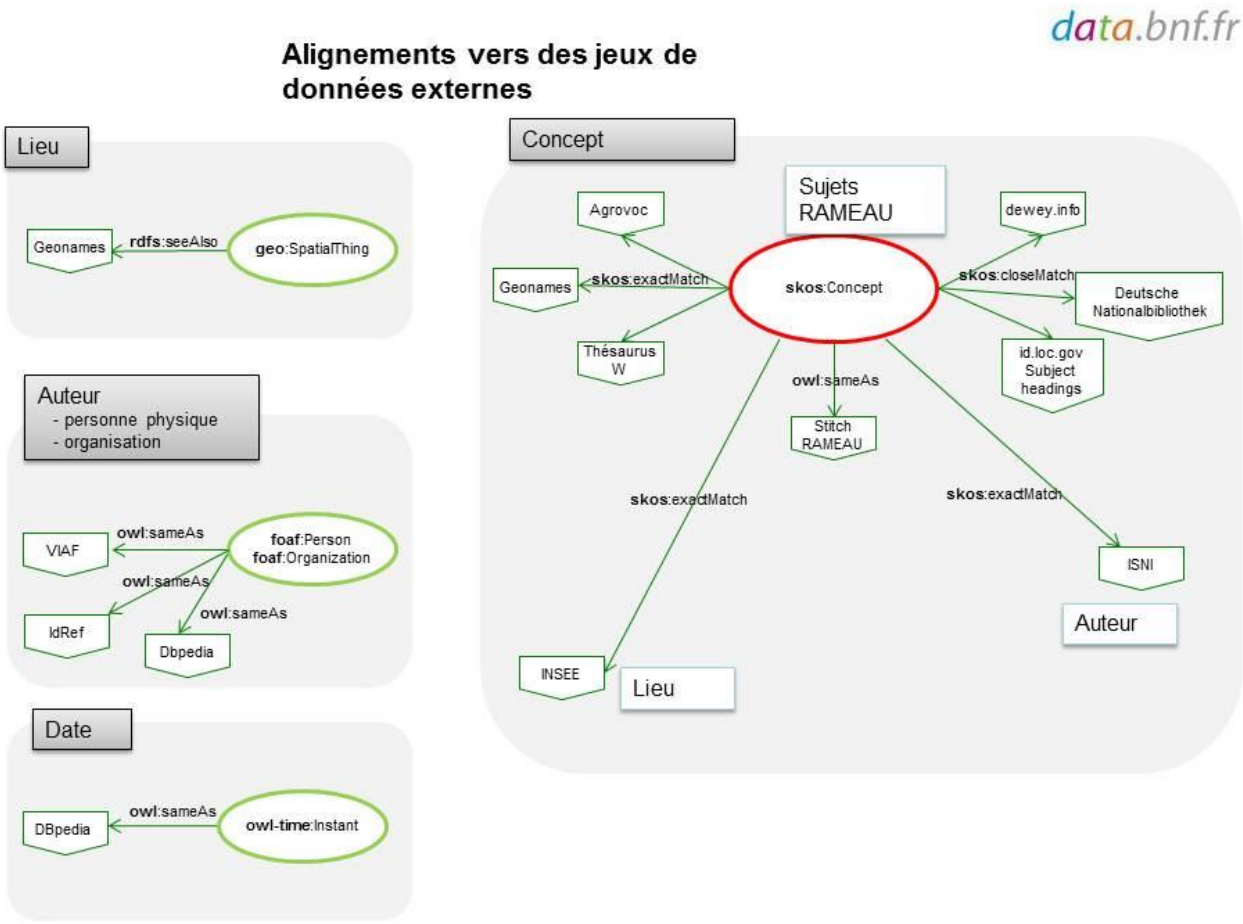
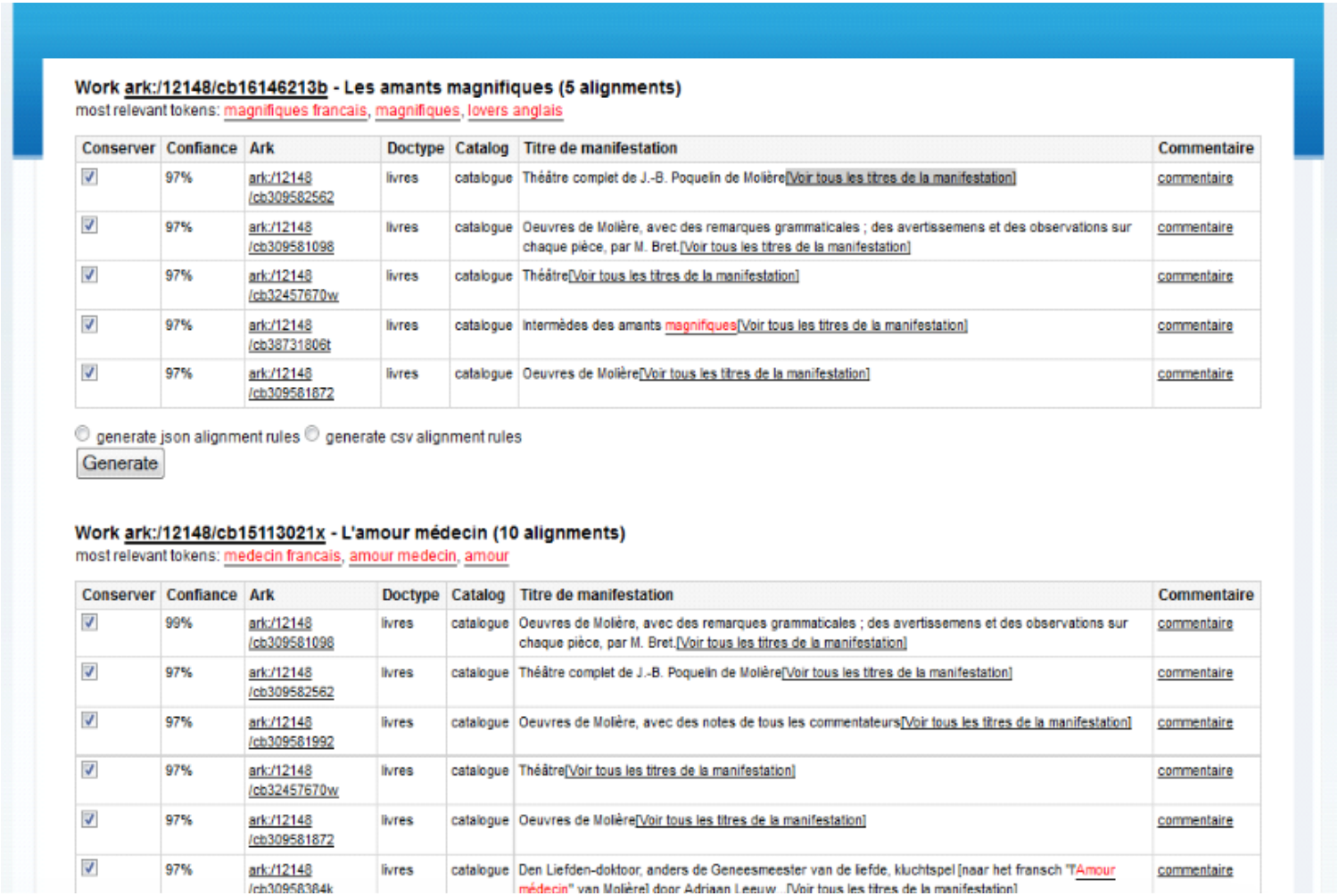


Figure 22. Alignements de data.bnf.fr vers des jeux de données externes

Comme mentionné précédemment à la section 5.1.1.4, les liens vers l'externe sont effectués grâce au logiciel Nazca développé par Logilab. Lorsque nécessaire, pour procéder à l'alignement, un outil a donc été développé visant à permettre à des correcteurs et des catalogueurs de la BnF de contrevérifier les suggestions offertes par la machine. La figure 23 montre les résultats présentés par l'outil pour les œuvres (au sens de l'entité du modèle FRBR)

de Molière (Simon, Di Mascio, Michel et Peyrard, 2014). On constate que les liens entre les ressources se font de manière automatique, mais que l'intervention humaine est parfois nécessaire pour confirmer ou infirmer un lien possible.



Work ark:/12148/cb16146213b - Les amants magnifiques (5 alignments)

most relevant tokens: magnifiques francais, magnifiques, lovers anglais

Conserver	Confiance	Ark	Doctype	Catalog	Titre de manifestation	Commentaire
☑	97%	ark:/12148/cb309582562	livres	catalogue	Théâtre complet de J.-B. Poquelin de Molière [Voir tous les titres de la manifestation]	commentaire
☑	97%	ark:/12148/cb309581098	livres	catalogue	Oeuvres de Molière, avec des remarques grammaticales ; des avertissemens et des observations sur chaque pièce, par M. Bret. [Voir tous les titres de la manifestation]	commentaire
☑	97%	ark:/12148/cb32457670w	livres	catalogue	Théâtre [Voir tous les titres de la manifestation]	commentaire
☑	97%	ark:/12148/cb38731806t	livres	catalogue	Intermèdes des amants magnifiques [Voir tous les titres de la manifestation]	commentaire
☑	97%	ark:/12148/cb309581872	livres	catalogue	Oeuvres de Molière [Voir tous les titres de la manifestation]	commentaire

○ generate json alignment rules ○ generate csv alignment rules

Generate

Work ark:/12148/cb15113021x - L'amour médecin (10 alignments)

most relevant tokens: medecin francais, amour medecin, amour

Conserver	Confiance	Ark	Doctype	Catalog	Titre de manifestation	Commentaire
☑	99%	ark:/12148/cb309581098	livres	catalogue	Oeuvres de Molière, avec des remarques grammaticales ; des avertissemens et des observations sur chaque pièce, par M. Bret. [Voir tous les titres de la manifestation]	commentaire
☑	97%	ark:/12148/cb309582562	livres	catalogue	Théâtre complet de J.-B. Poquelin de Molière [Voir tous les titres de la manifestation]	commentaire
☑	97%	ark:/12148/cb309581992	livres	catalogue	Oeuvres de Molière, avec des notes de tous les commentateurs [Voir tous les titres de la manifestation]	commentaire
☑	97%	ark:/12148/cb32457670w	livres	catalogue	Théâtre [Voir tous les titres de la manifestation]	commentaire
☑	97%	ark:/12148/cb309581872	livres	catalogue	Oeuvres de Molière [Voir tous les titres de la manifestation]	commentaire
☑	97%	ark:/12148/cb30958384k	livres	catalogue	Den Liefden-doktoor, anders de Geneesmeester van de liefde, kluchtspel [naar het fransch "l'Amour médecin" van Molière] door Adriaan Leeuw. [Voir tous les titres de la manifestation]	commentaire

Figure 23. Résultats présentés par l'outil de data.bnf.fr pour l'alignement des œuvres de Molière

6.1.10.2 Possibilités pour BAnQ

Pour BAnQ, il serait intéressant d'évaluer quelles autres instances travaillent sur des projets de Web de données au Québec et au Canada afin d'encourager la création de liens vers leurs ressources. Autrement, il est recommandé de lier les données vers des instances qui font office d'autorités dans leur domaine. Il serait donc possible pour l'institution d'effectuer des liens vers des institutions semblables, comme d'autres bibliothèques nationales qui participent au Web de données telles que LC, la BnF, la Bibliothèque nationale allemande, la Bibliothèque nationale d'Espagne, etc. Les liens vers les bases de données comme DBpedia, VIAF et Geonames sont aussi incontournables étant donné la place qu'elles occupent dans le Web de données.

6.1.11 Convertir les données en RDF

Cette étape ne pose en général pas trop de difficultés. Il est cependant nécessaire de choisir le ou les syntaxe(s) de sérialisation (voir section 2.5.2) dans lesquelles les données seront rendues disponibles pour les usagers qui voudraient éventuellement les réutiliser, mais surtout pour les machines. Il est important de savoir que la procédure doit se faire de façon automatique ou semi-automatique étant donné que d'effectuer un tel travail manuellement serait beaucoup trop fastidieux. Il existe un bon nombre d'outils qui permettent la conversion tels que Catmandu[46], D2RQ[47],OpenLink Virtuoso[48], etc.

[46] http://librecat.org/

[47] http://d2rq.org/

[48] http://virtuoso.openlinksw.com/

6.1.11.1 Le cas de data.bnf.fr

Pour sa part, le projet data.bnf.fr rend ses données RDF disponibles sous les formats suivants : RDF/XML, N-Triples, N3, JSON-LD et JSON. La conversion se fait via l'outil CubicWeb développé par son partenaire Logilab.

6.1.11.2 Possibilités pour BAnQ

L'utilisation d'un outil de conversion automatique est nécessaire et il est conseillé de rendre accessibles les données dans plusieurs syntaxes de sérialisation. Étant donné que tous les usagers ne travaillent pas avec le même format, il est préférable de procéder à plusieurs conversions que de perdre un usager potentiel qui souhaiterait procéder au téléchargement des données.

6.1.12 Valider les jeux de données

Avant de publier les jeux de données sur le Web, il est recommandé d'en vérifier la qualité. Bizer et Heath (2011) proposent une liste de points à vérifier avant de procéder à la publication. Le tableau 12 propose une liste de vérification.

Tableau 12. Points à vérifier avant de procéder à la publication des données sur le Web de données

Points à vérifier	☒
Est-ce que le jeu de données est lié à d'autres jeux de données ?	
Avez-vous rendu disponible la provenance des métadonnées ?	
Avez-vous rendu disponible une licence d'utilisation ?	
Utilisez-vous des termes provenant de vocabulaires ou ontologies bien établis ?	
Les URI sont-ils déréférençables ?	

Avez-vous effectué le *mapping* entre les termes de votre vocabulaire ou ontologie avec d'autres vocabulaires ou ontologies ?	
Avez-vous rendu disponible une description des jeux de données ?	

À cette étape, il est aussi possible de vérifier si les triplets RDF sont bien construits à l'aide d'outils de validation (Zengenene, Casarosa et Meghini, 2014) et de relire l'échelle de qualité basée sur cinq étoiles présentée à la section 2.4.

6.1.13 Publier les jeux de données

Une fois les données converties et évaluées, il est possible de les rendre accessibles sous forme de fichiers RDF (*dumps*), dans un *triplestore* présentant une interface d'interrogation SPARQL, via un point d'accès SPARQL, via un *RESTful API* (*Representational State Transfer*) ou encore directement dans les pages Web grâce à RDFa.

6.1.13.1 Le cas de data.bnf.fr

Comme mentionné précédemment, il est aussi possible pour les usagers de data.bnf.fr d'interroger les données via un point d'accès SPARQL. On peut aussi procéder au téléchargement de *dumps* ou au téléchargement d'une page unique dans les différents formats mentionnés ci-dessus. De plus, la BnF incorpore des données structurées internes dans ses pages Web.

6.1.13.2 Possibilités pour BAnQ

Ce choix dépendra de l'institution et des outils utilisés, mais en général on recommande de rendre les données disponibles en offrant le plus grand nombre d'options possibles afin de garantir et de faciliter les accès, autant du côté des usagers humains que des machines. Ainsi, il est recommandé de mettre sur pied un point d'accès SPARQL, mais aussi de rendre le

téléchargement des données disponibles. Les données structurées internes sont aussi une option intéressante et BAnQ met déjà ces technologies en pratique.

Notons aussi qu'il faut tenir compte de l'importance de la maintenance du système. Lorsque des changements sont effectués au catalogue ou à la base de données, il est nécessaire de faire des tests et des vérifications visant à s'assurer que ces modifications n'ont pas affecté le reste du système. De plus, il est à noter qu'un tel projet devra être toujours en évolution, principalement étant donné le fait que les technologies liées au Web sémantique sont développées en continu. Ainsi, il est possible qu'après une certaine période de temps, il soit nécessaire de réévaluer le travail effectué ainsi que l'architecture du système, dans l'objectif de pouvoir le simplifier ou en améliorer la performance, la stabilité et la fiabilité. De plus, de nouveaux acteurs peuvent prendre de l'importance et l'institution devra peut-être se retrouver en situation où elle devra créer de nouveaux liens de ses ressources vers ces acteurs. Il faut donc tenir compte de la nécessité de se tenir à jour sur ces technologies.

6.1.14 Tableau récapitulatif

Le tableau 13 présente un résumé des étapes présentées ci-dessus, des technologies et ressources nécessaires, des choix faits par la BnF ainsi que des possibilités qui s'offrent à BAnQ. Pour chaque étape présentée, on retrouve donc les considérations à prendre, c'est-à-dire les outils qui seront utilisés durant toute la phase de développement. Il permet de jeter un coup d'œil rapide sur les points développés précédemment. Il est aussi possible de se baser sur ce tableau afin de présenter de façon rapide les différentes étapes d'implantation du projet et de démontrer la faisabilité de celui-ci grâce aux choix effectués par la BnF.

Tableau 13. Résumé des étapes pour la mise sur pied d'un projet de Web de données, des technologies et ressources nécessaires, des choix effectués par la BnF et des possibilités pour BAnQ

Étapes	Technologies et ressources nécessaires	Le cas de data.bnf.fr	Possibilités pour BAnQ
Comprendre la motivation et prise de conscience	• Avantages du Web de données en bibliothèques		
Obtenir l'autorisation des parties prenantes	• Exemples de projets • Modèles représentant le cycle de vie des DOL • Liste des étapes de mise sur pied d'un tel projet		
Établir une licence d'utilisation	• *Creative Commons* • *ODbl* • Licence maison • Licence de l'état	• Licence Ouverte de l'État élaborée par Etalab	• Création de sa propre licence • Licence *Creative Commons* utilisée par le Gouvernement du Québec

Évaluer les compétences	• Connaissances en
	  ○ Systèmes d'information
	  ○ Métadonnées
	  ○ Modélisation
	• Communication
		• Équipe de six personnes à temps plein
		• Prestataire Logilab en charge du développement
			• Bibliothécaire de métadonnées
			• Analystes informatiques
			• Administrateur de base de données

Évaluer les jeux de données	• Notices se trouvant dans le catalogue ou la base de données • Référentiels ou notices d'autorité	• RAMEAU • Gallica • BNF Archives et manuscrits • BNF Catalogue général	• Notices d'autorité • Ressources patrimoniales • Données provenant de services • Données administratives

	• Éléments de métadonnées ou ontologies développées • Données administratives		
Choisir le modèle de publication	• À partir d'une base de données relationnelle • À partir d'un API • À partir d'un *triplestore* • À partir de données en langue naturelle	• À partir d'une base de données relationnelle	• À partir d'une base de données relationnelle • À partir d'un *triplestore*
Attribuer les URI	• URI multiples (concept, page HTML, flux RDF) • Identifiants opaques ou signifiants • Identifiants pérennes • Négociation de contenu	• Identifiants opaques • Identifiants pérennes (ARK) • URI multiples (concept, page HTML, flux RDF)	• Identifiants pérennes (ARK) générés par DSpace • Création d'URI nécessaire (au moins trois URI par concept)
Choisir son modèle de données et faire le *mapping*	• Choix des vocabulaires, ontologies et schémas de métadonnées • Construction d'ontologies si nécessaire • *Mapping*	• Utilisation de vocabulaires, ontologies et schémas de métadonnées multiples • Tableaux de conversion INTERMARC et EAD vers RDF	• *Mapping* EAD et MARC vers RDF • Choix de la structure sémantique doit être l'objet de discussions
Nettoyer les données	• Outils de nettoyage de données		• Utilisation d'un outil permettant le nettoyage automatique des données

Enrichir les données en faisant des liens	• Liens à l'interne
	• Liens à l'externe
		• Nazca (Logilab)
		• Liens vers l'externe
		  ○ Geonames
		  ○ IdRef
		  ○ DBpedia
			• Liens vers l'externe
			  ○ Autres bibliothèques nationales
			  ○ DBpedia
			  ○ Geonames

	• Utilisation de OWL et SKOS pour effectuer ces liens	○ LC ○ Bibliothèque nationale allemande ○ Thésaurus W ○ Agrovoc ○ Dewey ○ Insee ○ ISNI	
Convertir les données en RDF	• Choix de l'outil de conversion • Choix des ou de la syntaxe(s) de sérialisation	• CubicWeb (Logilab) • Syntaxes de sérialisation : RDF/XML, N-Triples, N3, JSON-LD et JSON	• Choix de l'outil de conversion • Choix des syntaxes de sérialisation (préférable d'offrir plusieurs options)
Valider les jeux de données	• Liste de vérification		
Publier les jeux de données	Plusieurs possibilités : • Point d'accès SPARQL • Téléchargement de *dumps* • Téléchargement de pages uniques • RESTful API • Données structures internes	• Point d'accès SPARQL • Téléchargement de *dumps* • Téléchargement de pages uniques • Données structurées internes	• Point d'accès SPARQL • Téléchargement de *dumps* • Données structurées internes
Maintenance du système	• Tests et vérifications suite à des modifications • Réévaluation de l'architecture du système • Création de nouveaux liens		

6.2 Un processus pour développement : le *Rational Unified Process*

Cette section a comme objectif de proposer une méthodologie de développement d'un logiciel pour l'application des étapes présentées à la section 6.1. Il est d'abord nécessaire de présenter les enjeux auxquels doivent faire face les professionnels de l'information lors de l'implémentation d'un projet informatique dans le cadre d'une institution documentaire.

6.2.1 Implémentation d'un projet informatique en bibliothèque

Les principes présentés ci-dessus se doivent d'être maîtrisés avant la mise sur pied d'un tel projet, c'est-à-dire que les professionnels de l'information qui y participeront se doivent d'être aptes à s'approprier ces technologies. Selon Jacquesson (1992), les bibliothèques n'utilisent que rarement les technologies informatiques les plus récentes et ce, pour les raisons suivantes :

- Les bibliothèques doivent assurer un service permanent, donc elles ne peuvent prendre de risques quant à l'utilisation d'outils informatiques qui ne sont pas nécessairement maîtrisés ;
- Les bibliothèques n'ont que rarement accès à un groupe de recherche et de développement qui permettrait d'effectuer des tests sur les nouveaux outils informatiques ;
- Les investissements qu'implique un changement de technologie informatique sont souvent trop élevés pour être mis en place au sein des bibliothèques ;
- La réticence au changement et la tendance conservatrice des professionnels de l'information.

Ainsi, ces difficultés peuvent faire en sorte que les projets impliquant de nouvelles technologies informatiques doivent souvent être présentés en chiffrant les bénéfices pour l'institution, ce qui peut être presque impossible dans certains cas. En effet, un projet tel que celui de la BnF présente souvent des résultats concrets à très long terme. Comme nous l'avons vu à la section 5.1.1.5, on arrive dans une certaine mesure à chiffrer les bénéfices (plus grand nombre de visiteurs sur les différentes plateformes, nombre de téléchargements de données et de projets de réutilisation de

celles-ci), mais il est nécessaire de mentionner que ce projet a maintenant cinq ans et que les rapports ont tardé à démontrer ce type de résultats. La plupart des conditions à un investissement nécessitent le fait de constater des résultats concrets et chiffrables, et ce, le plus rapidement possible. Ainsi, dans certains cas, cette difficulté peut freiner le développement de projets.

Le plus gros avantage que possèdent les bibliothèques en ce qui a trait au processus d'informatisation constant est le fait que leur capital est composé de données déjà structurées et à ce niveau, on pourrait dire que l'importance qui a été accordée à ce travail a été, en quelque sorte, visionnaire. Jacquesson (1992) souligne aussi le fait que le monopole du savoir et de la culture ne peut plus être entre les mains d'une seule entité avec l'arrivée et l'évolution de ces technologies. Ainsi, ce savoir et cette culture sont désormais l'objet d'un partage massif d'informations et un cadre informatique contrôlé au sein des institutions documentaires permettra d'avoir une meilleure vision d'ensemble de toute la masse d'informations accessibles au public. Pour ce faire, il indique qu'un changement au niveau des mentalités des gestionnaires de bibliothèque, une formation poussée des acteurs impliqués ainsi qu'une réflexion sur l'avenir du catalogue seront nécessaires (voir section 5.1.1.2). Cette analyse devrait d'abord prendre naissance au sein des bibliothèques nationales, qui font office d'autorités. Les attentes des usagers sont de plus en plus élevées quant à l'utilisation des nouvelles technologies dans les différents services utilisés et il en revient aux bibliothèques de s'adapter à ces nouveaux besoins.

6.2.2 *Rational Unified Process* (RUP)

Dans l'objectif de présenter un cadre quant aux différentes étapes nécessaires à la publication dans le Web de données, nous nous basons sur le processus de développement de logiciel itératif qu'est le RUP créé par Rational Software Corporation, qui appartient maintenant à la multinationale IBM (IBM, 2003-a). Sur la base de notre étude sommaire des processus possibles, celui-ci nous semblait le plus approprié à la situation décrite ici. Quelques processus de développement de logiciel ont été envisagés (*waterfall*, modèle en spirale, développement rapide d'applications, etc.) et le choix du processus itératif a été fait en fonction d'un tableau comparatif (voir tableau 14) proposé par Sami (2012).

Tableau 14. Tableau comparatif pregnant en compte différents processus de développement de logiciel et les facteurs à considerer pour fiare un choix parmi ceux-ci (Sami, 2012, notre traduction)

Facteurs à considérer	**Waterfall**	**Cycle en V**	**Prototypage évolutif**	**Modèle en spirale**	**Itératif et incrémental[49]**	**Méthodes agiles**
Exigences de l'utilisateur imprécises	Pauvre	Pauvre	Bon	Excellent	Bon	Excellent
Technologies inconnues	Pauvre	Pauvre	Excellent	Excellent	Bon	Pauvre
Système complexe	Bon	Bon	Excellent	Excellent	Bon	Pauvre
Système fiable	Bon	Bon	Pauvre	Excellent	Bon	Bon
Peu de temps dédié au développement	Pauvre	Pauvre	Bon	Excellent	Excellent	Excellent
Gestion de projet solide	Excellent	Excellent	Excellent	Excellent	Excellent	Excellent
Limitation des coûts	Pauvre	Pauvre	Pauvre	Pauvre	Excellent	Excellent
Visibilité des parties prenantes	Bon	Bon	Excellent	Excellent	Bon	Excellent
Compétences limitées	Bon	Bon	Pauvre	Pauvre	Bon	Pauvre
Documentation disponible sur le processus	Excellent	Excellent	Bon	Bon	Excellent	Pauvre
Possibilité de réutiliser les composantes	Excellent	Excellent	Pauvre	Pauvre	Excellent	Pauvre

Les facteurs à considérer selon nous pour ce type de projet étaient les technologies inconnues (bon), le manque de temps dédié au développement (excellent), la limitation des coûts (excellent), la visibilité des parties prenantes (bon), les compétences limitées (bon) et la documentation disponible sur le processus (excellent). Ainsi, en fonction de ce tableau, le choix a été fait de sélectionner un processus de développement de logiciel itératif et incrémental. Une fois ce choix effectué, nous avons explorer sommairement les différents processus existant dans

[49] Le processus RUP entre dans cette catégorie.

cette catégorie et avons arrêté notre choix sur le RUP étant donné que ce modèle permettait des modifications par l'équipe de développement si nécessaire ainsi qu'une possibilité de ne sélectionner que partiellement certains des éléments du modèle en fonction des besoins.

Il est d'abord nécessaire de présenter une synthèse du RUP afin d'assurer une meilleure compréhension du tableau 15. Le RUP est basé sur six principes (Kruchten, 2000) :

- Le développement itératif ;
- La gestion des besoins ;
- L'architecture et l'utilisation de composants ;
- La modélisation et le langage de modélisation unifié (UML) ;
- La qualité du processus et du produit ;
- La gestion de la configuration et des changements.

D'abord, le développement itératif est supérieur à une approche séquentielle linéaire (telle que le modèle *waterfall*[50] par exemple) dans la mesure où il permet de réévaluer les besoins en cours de développement, d'intégrer de nouveaux éléments au fur et à mesure, un meilleur contrôle des risques et la correction d'erreurs permettent d'obtenir une architecture plus robuste en fin de projet. De plus, tous les participants au projet peuvent travailler simultanément, donc ils apprennent plus rapidement les uns des autres (Krutchen, 2000; Munassar et Govardhan, 2010).

Ensuite, la gestion des besoins est une « (…) approche systématique visant à déterminer, organiser et gérer un cahier des charges toujours changeant d'un système informatique ou d'une application logicielle. » (Kruchten, 2000) Une gestion des besoins réussie permet donc un meilleur contrôle, une qualité améliorée, une diminution du temps et des coûts requis ainsi qu'une meilleure communication au sein de l'équipe de développement. Troisièmement, le RUP est orienté architecture dans la mesure où, dès le début du projet, on a comme objectif la conception d'une architecture fiable qui évoluera au fil des itérations et qui se base sur des cas

[50] Le modèle *waterfall* implique qu'une phase doit être complétée dans son entièreté avant que la prochaine phase puisse débuter.

d'utilisation (*use case*). Un cas d'utilisation permet de définir comment le système doit être utilisé, comment il doit interagir avec l'usager et quelles en sont les exigences fonctionnelles (Bittner et Spence, 2002).

Puis, le RUP accorde une grande importance à la modélisation afin de faciliter la compréhension des problèmes par la représentation et propose l'utilisation du langage de modélisation unifié (UML) pour exprimer les modèles. L'UML est un langage de modélisation graphique visant à standardiser la façon de visualiser la conception d'un système informatique (UML, 2005). La cinquième pratique recommandée est le fait que tous les membres de l'équipe de développement sont en charge de la qualité du produit en cours de production, mais aussi de la qualité du processus mis en place pour arriver au résultat final. Finalement, la gestion de la configuration et des changements implique le contrôle et le fait de garder les traces de tous les changements et erreurs qui ont été constatés durant les différentes phases de développement. Cela permet de suivre de façon plus rigoureuse le développement et l'avancement du projet (Kruchten, 2000).

La figure 24 illustre la structure d'un projet basé sur le RUP (IBM, 2003-b). La structure du RUP est basée sur deux dimensions : les phases de développement (axe horizontal sur la figure 24) et les activités qui s'y rattachent (axe vertical sur la figure 24). Les quatre phases de développement sont les suivantes : le commencement (*inception*), l'élaboration (*elaboration*), la construction (*construction*) et la transition (*transition*). Pour ce qui est des activités, on définit deux types. D'un côté, on retrouve les activités de base (*disciplines* dans la figure 24), soit la création du modèle d'affaires (*business modelling*), la définition des exigences (*requirements*), l'analyse et la conception (*analysis and design*), l'implémentation (*implementation*), les tests (*test*) et le déploiement (*deployment*). De l'autre côté, on retrouve les activités de support aux activités de base, soit la gestion de configuration logicielle (*configuration and change management*), la gestion de projet (*project management*) et la mise en place de l'environnement de développement (*environment*). Chacune des phases est caractérisée par des itérations (*Initial*, *Elab #1*, *Elab #2*, *Const #1, Const #2, Const #N, Tran #1* et *Tran #2* dans la figure 24). Cela

permet de diminuer les chances de risques qui peuvent être plus fréquents lorsque le travail n'est pas testé au fur et à mesure qu'il est développé.

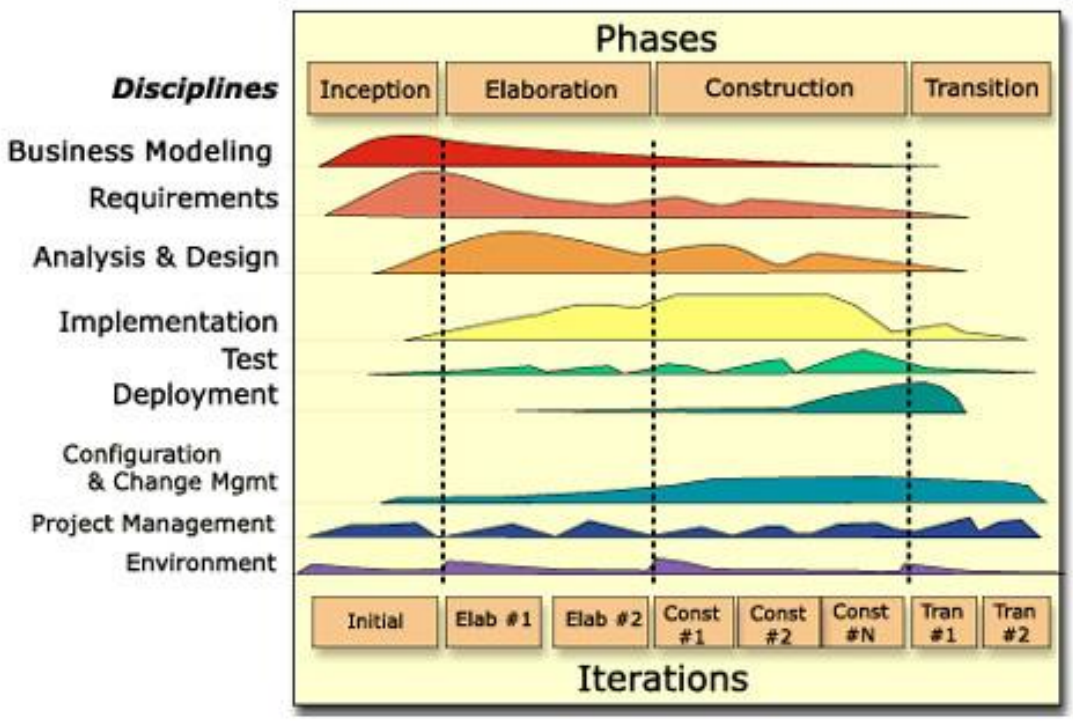


Figure 24. Structure d'un projet basé sur le RUP

Le commencement (phase 1) se caractérise par la définition de la portée du projet en évaluant les coûts potentiels, le budget ainsi que les facteurs de succès. On doit aussi mettre sur papier un cas d'utilisation, une évaluation des ressources nécessaires, un rapport d'évaluation des risques en ce qui a trait à la faisabilité, un plan de projet ainsi qu'une description de celui-ci. L'objectif de cette première phase est aussi de proposer une architecture du système. Comme mentionné précédemment, chacune des phases peut être retravaillée, corrigée et recommencée tant que l'équipe n'en est pas satisfaite (Rational Software, 1998). Ceci s'applique à toutes les phases de développement. Ensuite, la phase d'élaboration (phase 2), que certains considèrent comme la plus importante (Kruchten, 2000; Rational Software, 1998), permet d'assurer que l'architecture, les exigences, le coût, les délais requis et les risques ont été analysés en profondeur grâce aux versions presque définitives des cas d'utilisation. Il s'agit de la dernière phase avant le début de la constitution réelle du projet. Ensuite, la phase de construction (phase

3) permet de finaliser les cas d'utilisation, mais aussi de construire toute l'architecture et de travailler aux aspects plus techniques de développement. Finalement, la dernière phase, la transition (phase 4), permet d'effectuer le déploiement du projet, la production de la documentation sur le sujet et la correction finale d'erreurs. Une fois toutes ces phases terminées, le processus de développement prend fin du même coup.

Ensuite, en ce qui a trait aux activités, nous aborderons uniquement les activités de base, soit la création du modèle d'entreprise, la définition des exigences, l'analyse et la conception, l'implémentation, les tests et le déploiement. D'abord, la création du modèle d'entreprise nécessite une excellente communication entre les différentes parties prenantes (Rational Software, 1998). Les cas d'utilisation sont mis de l'avant et permettent la création d'unités de fonctionnalité du système, d'unités de développement, d'unités de tests et d'unités de réutilisations. Ces différents blocs visent à analyser la qualité de l'architecture et à développer une documentation claire présentant les ressources nécessaires à l'implémentation du projet. Ensuite, la définition des exigences a comme objectif d'identifier ce que le système doit faire (*quoi?*), à l'aide des cas d'utilisation (Rational Software, 1998). L'analyse et la conception, à partir des exigences, visent à décrire les spécificités du système (*comment?*) et à proposer un plan de conception. Ensuite, l'implémentation vise à mettre sur pied l'architecture grâce aux outils et aux compétences techniques. Les tests, une activité primordiale, doivent être prévus, conçus, effectués et évalués avant la phase de déploiement. Finalement, le déploiement implique la publication, la distribution, l'installation et l'aide aux usagers (Rational Software, 1998). En basant un tel projet sur le RUP, il serait possible, selon les auteurs, de diminuer les risques, de mieux gérer les coûts et les délais relatifs au développement, de travailler de manière itérative et de développer une meilleure communication au sein de l'équipe de développement.

Il est possible d'évaluer les différentes étapes présentées à la section 6.1 ci-dessus et d'y appliquer les phases ainsi que les activités de base du RUP. Le tableau 15 résume cette analyse. Chacune des étapes doit être travaillée en suivant les phases du RUP, mais en procédant de manière itérative. Chaque membre de l'équipe de développement est impliqué dans toutes les étapes et la communication au sein de l'équipe est grandement encouragée.

Tableau 15. Phases et activités de base principales du RUP appliquées aux étapes présentées à la section 6.1

Phases / Activités de base	Commencement	Élaboration	Construction	Transition
Création du modèle d'entreprise	Comprendre la motivation et prise de conscience Obtenir l'autorisation des parties prenantes	Établir une licence d'utilisation		
Définition des exigences		Évaluer les compétences Évaluer les jeux de données Choisir le modèle de publication et évaluer les outils nécessaires		
Analyse et conception			Attribuer les URI Choisir son modèle de données et faire le *mapping*	
Implémentation			Nettoyer les données Enrichir les données en faisant des liens Convertir les données en RDF	
Tests				Valider les jeux de données
Déploiement				Publier les jeux de données

6.3 Recommandations et défis pour BAnQ

Mis à part le fait de suivre les étapes identifiées ci-dessus, BAnQ devra, dans la mesure où la décision d'implanter un projet de Web de données est prise, faire face à des défis particuliers. Nous invitons le lecteur à se référer au tableau 13 présenté à la section 6.1.14 qui présente un résumé des technologies et ressources nécessaires pour une telle implantation. Il est aussi possible de se baser sur le tableau 15 présenté à la section 6.2.2 pour évaluer les phases et activités du RUP qui s'appliquent aux étapes.

D'abord, les deux premières étapes, soit « Comprendre la motivation et prise de conscience » (section 6.1.1) et « Obtenir l'autorisation des parties prenantes » (section 6.1.2) font partie de la première phase du RUP (la phase de commencement). Ces deux étapes misent sur les activités de création du modèle d'affaires et touchent aussi à la définition des exigences.

Ensuite, l'étape du choix de la licence d'utilisation fait partie de la phase d'élaboration (phase 2) dans le RUP. Elle répond aux objectifs de cette phase, soit d'évaluer ce que le système peut permettre et jusqu'à quel niveau. Ainsi, en analysant les cas d'utilisation (au sens RUP), il est possible de rédiger une licence qui correspondra à la fois aux besoins des usagers potentiels et aux besoins de l'institution. De plus, comme indiqué à la section 6.2.2, le RUP souligne l'importance de la participation active des différents membres de l'équipe à toutes les phases de développement ainsi que la nécessité d'assurer une excellente communication entre eux. Les étapes d'évaluation des compétences, d'évaluation des jeux de données ainsi que du choix du modèle de publication font aussi partie de la phase d'élaboration du RUP et permettent d'identifier les exigences pour la phase suivante qui consiste à mettre sur pied l'architecture du système. Ces analyses sont donc primordiales aux activités de base d'analyse et de conception, de définition des exigences et d'implémentation. Elles permettent d'évaluer les risques reliés à chaque mode de publication ainsi que la stabilité pour chacun. On commence alors à concevoir de façon plus claire l'architecture.

Ensuite, la phase de construction débute par la pratique d'attribution des URI. Il s'agit du premier travail de construction nécessitant des connaissances techniques ainsi qu'une compréhension approfondie du contenu de la base de données. Le choix du modèle de données et l'activité de *mapping* font aussi partie de cette phase et découlent de l'activité d'analyse et de conception. Il s'agit d'un travail important sur lequel les membres de l'équipe devront inévitablement revenir à plusieurs reprises afin de s'assurer que tous les cas de figure auront été analysés. Puis, les étapes de nettoyage de données, d'enrichissement des données à l'aide de liens et de conversion de celles-ci vers RDF découlent de l'activité d'implémentation. Il s'agit d'étapes cruciales d'entrée dans le Web de données et elles permettent d'organiser les ressources, de continuer à la construction de l'architecture du système et d'augmenter les possibilités de réutilisation.

L'étape de validation des jeux de données, pour sa part, est la première étape faisant partie de la dernière phase de développement, soit la phase de transition. Elle se fait par le biais de l'activité de tests. Notons que des tests doivent tout de même être effectués pour chaque itération. Finalement, la dernière étape de publication des données est caractérisée par les activités de déploiement, soit de finalisation du projet et de tests finaux avant la publication.

Le Web de données et le Web sémantique étant des concepts pouvant être abstraits pour certains, il peut être difficile de les présenter aux instances. Certains aspects étant très techniques, il faut présenter ces technologies de manière à ce que tous les acteurs impliqués puissent se les approprier et en comprendre la portée. Surtout, il est primordial d'aborder la question sous l'angle stratégique pour l'institution, car exercer un leadership en matière de Web sémantique implique un travail à haut niveau qui permettra de susciter l'engagement et l'adhésion. Une discussion quant au futur du catalogue de bibliothèque sera éventuellement inévitable et il faut noter que plusieurs départements seront alors appelés à communiquer et collaborer. Ainsi, la proposition d'utiliser le processus RUP est à envisager. Ce processus mise beaucoup sur la communication et sur l'évaluation des risques. En travaillant à l'aide de cette ligne directrice et en portant une importance particulière à la documentation tout au long du

processus de développement, il sera plus facile d'évaluer les coûts et les délais requis pour la mise en place d'un tel projet, mais aussi de permettre la présentation du projet par la suite. En sa qualité de bibliothèque nationale, BAnQ pourrait alors jouer un rôle important de formateur pour d'autres institutions documentaires québécoises qui souhaiteraient implémenter en son sein un tel projet par la suite. Il est fortement encouragé de produire cette éventuelle documentation, car plus il y a d'acteurs qui participent au Web de données, plus les liens entre les ressources sont possibles, plus la visibilité des collections est augmentée sur le Web et plus les possibilités de réutilisation sont nombreuses.

Ensuite, il est à prévoir qu'une certaine infrastructure informatique sera nécessaire. Ainsi, il est conseillé de travailler le plus possible avec des logiciels libres et gratuits afin de diminuer les coûts de développement et d'être en phase avec les stratégies gouvernementales (voir section 6.1.6.2). Une recherche plus approfondie sera peut-être nécessaire afin d'effectuer un choix relatif aux options technologiques en fonction des caractéristiques et des besoins de l'institution. Le développement à l'interne est recommandé, car l'emploi de ressources externes peut aussi en général impliquer un coût élevé étant donné l'expertise nécessaire. De plus, BAnQ travaillant avec le SIGB Portfolio[51], on constate une certaine dépendance à cette technologie et à son évolution. En effet, étant donné que Portfolio ne propose pas de solutions technologiques en lien avec le Web sémantique, si BAnQ souhaite mettre sur un pied un projet déployant ce type de technologies, l'institution doit élaborer des outils externes au SIGB tout en assurant une certaine interopérabilité.

Un autre défi est celui de la modélisation des données. Au niveau de la conceptualisation, certaines difficultés peuvent se présenter et il peut être ardu pour tous les membres de l'équipe de s'entendre sur le processus de *mapping* ainsi que sur le choix des ontologies, vocabulaires et schémas de métadonnées qui pourront décrire adéquatement les ressources numériques que l'institution a en sa possession. Aucune ontologie n'a encore été créée pour répondre parfaitement aux besoins d'une bibliothèque nationale dont la collection comporte aussi un

[51] http://www.bibliomondofr.com/#!pour-bibliothques/c1gwy

grand nombre d'archives et il n'y a aucun consensus dans le domaine quant au meilleur choix. Il faut donc prévoir l'utilisation de plusieurs outils pour arriver à une description qui conviendra à tous et accepter de perdre un certain niveau de description des ressources. Il est donc recommandé d'analyser les différentes options qui s'offrent à BAnQ tout en se basant sur des cas d'utilisation[52]. Un travail de création d'ontologie serait trop coûteux, il est donc préférable de se baser sur des ontologies et vocabulaires déjà éprouvés par d'autres institutions semblables.

Mentionnons finalement qu'il serait possible pour BAnQ de profiter des partenariats déjà établis avec d'autres institutions afin de commencer par un projet-pilote qui lui permettrait de se familiariser avec les technologies du Web sémantique. Ainsi, les Archives Canada-France (BAC) ou encore le RFN (BnF) seraient deux exemples de projets déjà actifs qui pourraient servir d'essais pour la publication de données ouvertes et liées.

[52] Un exemple de cas d'utilisation pourrait être celui d'un usager qui voudrait accéder aux données liées aux événements qui ont lieu à la Grande bibliothèque pour pouvoir créer une application présentant les différentes activités gratuites offertes au moins de 10 ans à Montréal. Ainsi, il faudrait analyser les différentes étapes pour l'usager ainsi que les réactions du système.

7. Conclusion

L'objectif général de ce mémoire était d'explorer un ensemble de technologies spécifiques au Web sémantique et au Web de données, puis de proposer une liste de treize étapes visant l'utilisation et l'appropriation de ces technologies par les professionnels de l'information. En se basant sur une revue de la littérature approfondie, il fut possible d'identifier ces étapes pour le développement d'un projet de Web de données et de les appliquer par la suite au travail innovateur effectué par la Bibliothèque nationale de France dans le cadre du projet data.bnf.fr. Le travail de synthèse relatif à la définition des concepts et des technologies qui composent le Web sémantique et le Web de données permet de faciliter la compréhension de ces étapes. Puis, grâce à la méthodologie RUP présentée à la section 6.2.2, il est possible de proposer des recommandations pour BAnQ dans l'éventualité où un projet de Web de données prendrait forme dans cette institution.

On constate que le Web de données permet de jeter un nouveau regard sur les occasions d'innover qui se présentent aux professionnels de l'information. Les comportements informationnels ayant grandement changé depuis l'arrivée du Web et des ressources numériques ainsi que l'évolution vers une société de la connaissance, les défis sont nombreux et les bibliothèques et institutions documentaires doivent s'adapter tout en travaillant au processus d'évolution du catalogue. Le modèle FRBR, présenté à la section 5.1.1.2, est un exemple d'outil qui pourrait avoir une incidence sur la gestion de l'information de par sa façon de conceptualiser la notice bibliographique.

Comme nous l'avons présenté à la section 6.1.8, le processus de modélisation et de *mapping* nécessaire avant la conversion des données vers le format RDF peut présenter des difficultés, impliquer une perte de granularité de l'information et monopoliser des ressources. En 2012, la Bibliothèque du Congrès a annoncé l'embauche d'une firme externe nommée Zepheira afin que celle-ci travaille au développement d'un nouveau modèle de données pour la description bibliographique (Library of Congress, 2012-a). Ce modèle fut nommé le

Bibliographic Framework[53] (BIBFRAME) et a été développé dans l'objectif clair de remplacer les formats MARC et de fournir un format qui serait compatible avec le Web de données (Library of Congress, 2012-b). Le modèle BIBFRAME (maintenant appelé BIBFRAME 2.0 depuis le 21 avril 2016), est un travail en constante évolution et est basé sur trois niveaux d'abstraction (Library of Congress, 2016) :

- Œuvre (*work*);
    - Reflète l'essence conceptuelle de la ressource à cataloguer.
- Instance (*instance*);
    - Une œuvre peut avoir une ou plusieurs réalisation(s) matérielle(s). Ces réalisations sont les instances de l'œuvre.
- Item (*item*).
    - Un item et une copie physique ou électronique de l'instance.

Le modèle BIBFRAME n'est donc pas sans rappeler le modèle FRBR[54]. Ainsi, l'œuvre présente des informations telles que les auteurs, la langue ou les sujets. L'instance, pour sa part, donne de l'information sur l'éditeur, la date de publication, le format, etc. Finalement, l'item présente des données quant à sa location, son code à barres, etc. Ensuite, BIBFRAME propose d'autres concepts qui entrent en relation avec les trois niveaux d'abstraction (Library of Congress, 2016), soit les :

- Agents (*agents*);
    - Les personnes ou les organisations associées à une œuvre ou une instance en raison de leurs rôles en tant qu'auteur, éditeur, artiste, photographe, compositeur, illustrateur, etc.
- Sujets (*subjects*);

[53] http://bibframe.org/

[54] Selon la Library of Congress (2012-b), les niveaux Œuvre et Instance sont effectivement semblables aux relations trouvées dans le modèle FRBR. Cependant, BIBFRAME permet d'illustrer ces relations en graphe et non de façon hiérarchique. Cette façon de faire a comme objectif de simplifier la conceptualisation et la compréhension du modèle.

- Une œuvre peut être à propos d'un ou plusieurs concept(s). Ces concepts sont considérés comme les sujets de l'œuvre. Il peut s'agir de thèmes, lieux, événements, d'autres agents, d'autres items, etc.

- Événements (*events*).
  - Un événement est quelque chose qui a lieu à un endroit spécifique pendant un temps spécifique.

La figure 25 (Library of Congress, 2016) permet d'illustrer les relations entre les différents concepts présentés ci-dessus.

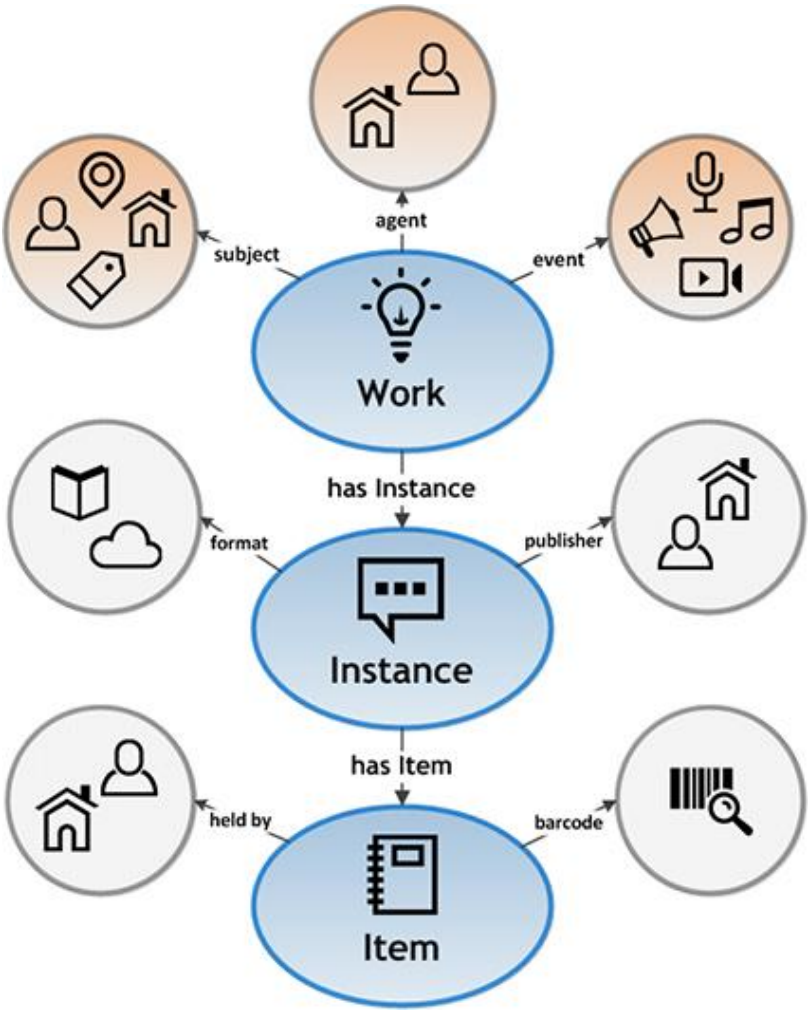


Figure 25. Modèle BIBFRAME 2.0

De plus, le modèle BIBFRAME 2.0 propose un vocabulaire qui est constitué de classes et de propriétés. Les classes incluent les trois niveaux d'abstraction ainsi que d'autres classes additionnelles. Les propriétés décrivent les caractéristiques relatives à la ressource et les relations entre les ressources. L'objectif ici n'est pas de décrire de façon détaillée le modèle BIBFRAME 2.0, mais plutôt d'accorder une importance au fait que la Bibliothèque du Congrès travaille de façon active à développer un modèle qui répondrait aux besoins liés au Web de données pour les bibliothèques. Ce modèle est basé sur RDF et sur les principes du Web sémantique. Dans la mesure où un tel modèle était éventuellement adopté par toutes les institutions documentaires, on constaterait une plus grande facilité à travailler avec ces nouvelles technologies. Le format MARC n'est malheureusement pas un format qui est compatible avec les nouvelles technologies auxquelles les professionnels de l'information doivent maintenant faire face. De plus, BIBFRAME 2.0 propose la diffusion sur le Web des métadonnées de bibliothèques et, ainsi, une meilleure visibilité (Hawkins, 2015). Cependant, la transition vers ce modèle n'est pas imminente et celui-ci est toujours à l'étude. Par contre, on constate un intérêt de plus en plus marqué pour celui-ci et pour les technologies du Web sémantique. BIBFRAME 2.0 pourrait donc se présenter éventuellement comme une solution à l'implémentation de projet de Web de données dans les institutions documentaires, par le fait que le modèle a été construit expressément pour répondre aux besoins des bibliothèques.

Ce nouvel environnement potentiel pour les institutions documentaires présente de nombreux avantages qui ont été présentés au chapitre 3 et la documentation sur le sujet est de plus en plus présente. Cependant, on ressent tout de même un certain ralentissement du côté de la mise sur pied de projets mettant en œuvre les technologies du Web sémantique. En effet, les projets documentés datent souvent de quelques années et bien que ceux-ci soient toujours maintenus, on ressent une certaine gêne du côté de la francophonie, la BnF étant la seule institution nationale francophone s'étant lancée dans un tel projet. Ainsi, il est pertinent de se questionner à savoir si les bibliothèques attendent que le modèle BIBFRAME ainsi que les autres technologies du Web de données soient assez évolués et ont été assez testés avant de tenter de les implanter. Un tel changement implique de nombreuses modifications quant aux

habitudes de travail et un temps non négligeable quant à la formation nécessaire pour mettre en pratique ces connaissances. Ainsi, les ressources nécessaires sont importantes et une réflexion s'impose au niveau des décideurs et des instances gouvernementales. On constate au Québec un certain intérêt à travers le Plan culturel numérique (section 5.2) et il sera intéressant de suivre l'évolution des projets mis en branle dans le cadre de cette initiative. Dans tous les cas, il sera éventuellement nécessaire pour BAnQ de se positionner quant aux technologies du Web sémantique et les prochaines années devraient voir naître de tels projets.

Il est permis de croire que les projets de Web de données au sein d'institutions documentaires prendront de plus en plus d'ampleur lorsque les pratiques auront été établies de manière claire. De plus, l'arrivée de BIBFRAME 2.0 joue inévitablement un rôle important dans la mesure où elle permet d'envisager l'encodage des métadonnées de bibliothèque de manière à ce qu'elle soit déjà prête à la publication dans le Web sémantique. Dans tous les cas, l'exemple de la BnF ainsi que d'autres démontrent que le Web de données en bibliothèque peut fonctionner et est profitable, mais que cela nécessite un travail évolutif et constant. Du côté de BAnQ, il semble plus que pertinent pour l'institution d'envisager la mise sur pied d'un tel projet, principalement dû au fait qu'elle fait office d'autorité et qu'elle peut devenir un modèle pour d'autres institutions, mais aussi parce qu'elle pourra ainsi contribuer à l'économie de la culture québécoise. L'intérêt pour le Web sémantique se fait tranquillement sentir au Québec et il s'agit d'une occasion pour BAnQ d'innover, de faire partie de ce nouveau mouvement international, évolutif et collaboratif ainsi que de s'inscrire dans un projet où l'objectif est d'améliorer les services aux usagers.

Références bibliographiques

Aharony, N. (2009). The influence of LIS students' personality characteristics on their perceptions towards Web 2.0 use. *Journal of Librarianship and Information Science*, *41*(4), 227-242.

Alemu, G., Stevens, B., Ross, P. et Chandler, J. (2012). Linked data for libraries : Benefits of a conceptual shift from library-specific record structures to RDF-based data models. *New Library World*, *113*(11/12), 549-570.

Baker, T., Bermès, E., Coyle, K., Dunsire, G., Isaac, A., Murray, P., … Zeng, M. (2011). Library linked data incubator group final report : W3C incubator group report 25 October 2011. Repéré à http://www.w3.org/2005/Incubator/lld/XGR-lld/

BAC. (2016). Archives de la Nouvelle-France - bibliothèque et archives Canada. Repéré à http://www.bac-lac.gc.ca/fra/decouvrez/exploration-colonisation/archives-nouvelle-france/Pages/archives-nouvelle-france.aspx

BAnQ. (2015-a). Rapport annuel de gestion 2014-2015. Repéré à http://www.banq.qc.ca/documents/a_propos_banq/rapports_annuels/BAnQ_RapportAnnuelDeGestion_2014-2015_br.pdf

BAnQ. (2015-b). Vision et mission | BAnQ numérique. Repéré à http://numerique.banq.qc.ca/p/apropos/vision_mission.html

BAnQ. (s.d.-a). Historique. Repéré à http://www.banq.qc.ca/a_propos_banq/historique/

BAnQ. (s.d.-b). Mission. Repéré à http://www.banq.qc.ca/a_propos_banq/mission_lois_reglements/mission/

BAnQ. (s.d.-c). Organisation. Repéré à http://www.banq.qc.ca/a_propos_banq/organisation/

BAnQ. (s.d.-d). Collection numérique. Repéré à http://www.banq.qc.ca/collections/collection_numerique/index.html

BAnQ. (s.d.-e). Comment chercher dans les archives. Repéré à http://www.banq.qc.ca/archives/entrez_archives/comment_chercher_archives/

BAnQ. (s.d.-f). Prix littéraires du Québec. Repéré à http://services.banq.qc.ca/sdx/prix/accueil.xsp?db=prix_litteraire

BAnQ. (s.d.-g). À propos du Catalogue des bibliothèques du Québec. Repéré à http://cbq.banq.qc.ca/cbq/f?p=404:31:::NO:::

Banque mondiale. (2016). *World development report 2016 : Digital dividends*. Washington, DC : World Bank. doi:10.1596/978-1-4648-0671-1

Barnaghi, P., Wang, W., Henson, C. et Taylor, K. (2012). Semantics of the internet of things : Early progress and back to the future. *International Journal on Semantic Web & Information Systems*, *8*(1), 1-21.

Bennett, R., Lavoie, B. F. et O'Neill, E. T. (2013). The concept of a work in WorldCat : An application of FRBR. *Library Collections, Acquisitions, and Technical Services*, *27*(1), 45-59.

Bermès, E. (2013). Enabling your catalogue for the semantic web. Dans S. Chambers (dir.), *Catalogue 2.0 : the future of the library catalogue* (p. 117-142). Chicago : Neal-Schuman.

Bermès, E. et Fauduet, L. (2011). The human face of digital preservation : Organizational and staff challenges, and initiatives at the Bibliothèque nationale de France. *The International Journal of Digital Curation*, *6*(1), 226-237.

Bermès, E., Isaac, A. et Poupeau, G. (2013). *Le web sémantique en bibliothèque*. Paris : Electre/Éditions du Cercle de la Librairie.

Berners-Lee, T. (1998). Cool URIs don't change. Repéré à http://www.w3.org/Provider/Style/URI.html

Berners-Lee, T. (2006). Linked data. Repéré à http://www.w3.org/DesignIssues/LinkedData.html

Berners-Lee, T. (2009, février). *Tim Berners-Lee : the next web* [Vidéo en ligne]. Repéré à https://www.ted.com/talks/tim_berners_lee_on_the_next_web

Berners-Lee, T. (2010). Is your linked open data 5 star? Repéré à https://www.w3.org/DesignIssues/LinkedData.html

Berners-Lee, T., Cailliau, R., Luotonen, A., Nielsen, H. F. et Secret. (1994). The world-wide web. *Communications of the ACM*, *37*(8), 76-82. doi: 10.1145/179606.179671

Berners-Lee, T., Fielding, R. et Masinter, L. (2005). Uniform resource identifier (URI) : Generic syntax (No. RFC 3986). Repéré à http://tools.ietf.org/html/rfc3986

Berners-Lee, T., Hendler, J. et Lassila, O. (2001). The semantic web. *Scientific American 284*(5), 28-37.

Bibliothèque numérique mondiale. (s.d.-a). Bibliothèque et Archives nationales du Québec. Repéré à https://www.wdl.org/fr/institution/#national-library-and-archives-quebec

Bibliothèque numérique mondiale. (s.d.-b). Bibliothèque nationale de France. Repéré à https://www.wdl.org/fr/institution/#national-library-of-france

Bibliothèque numérique mondiale. (s.d.-c). WDL descriptive metadata element set. Repéré à http://project.wdl.org/standards/metadata.html

Bittner, K. et Spence, I. (2002). *Use case modeling*. Boston : Addison-Wesley Professional.

Bizer, C., Cyganiak, R. et Heath, T. (2007). How to publish linked data on the web. Repéré à http://wifo5-03.informatik.uni-mannheim.de/bizer/pub/LinkedDataTutorial/

Bizer, C., Heath, T. et Berners-Lee, T. (2009). Linked data : The story so far. *Semantic Services, Interoperability and Web Applications : Emerging Concepts*, 205-227.

Bizer, C. et Heath, T. (2011). *Linked data : Evolving the web into a global data space*. San Rafael, California : Morgan & Claypool.

BnF. (s.d.-a). Mandragore, base des manuscrits enluminés de la BnF. Repéré à http://mandragore.bnf.fr/html/accueil.html

BnF. (s.d.-b). Standford prize for innovation in research libraries (SPIRL) – Application from the Bibliothèque nationale de France (BnF) for Gallica (gallica.bnf.fr) and Data (data.bnf.fr). Repéré à https://library.stanford.edu/sites/default/files/Bibliotheque%20nationale%20de%20France.pdf

BnF. (2013-a). Gallica et ses partenaires. Repéré à http://www.bnf.fr/documents/GALLICA_fiche1_partenaires.pdf

BnF. (2013-b). Les métadonnées de préservation numérique. Repéré à http://www.bnf.fr/fr/professionnels/preservation_numerique_boite_outils/a.pres_num_metadonnees.html

BnF. (2013-c). L'expérimentation OpenCat. Repéré à http://www.bnf.fr/fr/professionnels/web_donnees_applications_bnf/a.opencat.html

BnF. (2014-a). Organisation de la BnF. Repéré à http://www.bnf.fr/fr/la_bnf/missions_bnf/s.organisation_bnf.html?first_Art=non

BnF. (2014-b). BnF Catalogue général. Repéré à http://www.bnf.fr/fr/collections_et_services/catalogues_en_ligne/a.bnf_catalogue_general.html

BnF. (2014-c). La bibliothèque en chiffres. Repéré à http://emploi.bnf.fr/DRH/emploi.nsf/IXURL02/P2014000383_La-bibliotheque-en-chiffres

BnF. (2014-d). ReLIRE | Le cadre légal. Repéré à https://relire.bnf.fr/projet-relire-cadre-legal

BnF. (2015-a). Missions et projets de la BnF. Repéré à http://www.bnf.fr/fr/la_bnf/missions_bnf.html

BnF. (2015-b). Rapport d'activité 2014. Repéré à http://webapp.bnf.fr/rapport/pdf/rapport_2014.pdf

BnF. (2015-c). Gallica, la Bibliothèque numérique de la BnF et ses partenaires. Repéré à http://www.bnf.fr/fr/collections_et_services/bibliotheques_numeriques_gallica/a.gallica_bibliotheque_numerique_bnf.html#SHDC__Attribute_BlocArticle6BnF

BnF. (2015-d). Réseau national de coopération. Repéré à http://www.bnf.fr/fr/professionnels/cooperation_nationale/a.reseau_national_cooperation.htht

BnF. (2015-e). Gallica et la numérisation concertée. Repéré à http://www.bnf.fr/fr/professionnels/cooperation_nationale/a.gallica_numerisation_partagee.html

BnF. (2015-f). VIAF (Virtual International Authory File). Repéré à http://www.bnf.fr/fr/professionnels/donnees_autorites/a.viaf.html

BnF. (2015-g). À propos de data.bnf.fr. Repéré à http://data.bnf.fr/fr/about

BnF. (2015-h). Ouverture des données publiques. Repéré à http://www.bnf.fr/fr/professionnels/anx_recuperation_donnees/a.ouverture_donnees_bnf.html

BnF. (2016). Web sémantique et modèle de données. Repéré à http://data.bnf.fr/fr/semanticweb

Bradley, P. (2015). *Social media for creative libraries.* London : Facet Publishing.

Brickley, D. et Miller, L. (2014). FOAF vocabulary specification 0.99. Repéré à http://xmlns.com/foaf/spec/

Callewaert, R. (2013). FRBRizing your catalogue: the facets of FRBR. Dans S. Chambers (dir.), *Catalogue 2.0 : the future of the library catalogue* (p. 93-115). Chicago : Neal-Schuman.

Cantara, L. (2005). METS : the metadata encoding and transmission standard. *Cataloging & classification quarterly*, *40*(3-4), 2237-253. doi: 10.1300/J104v40n03_11

Caplan, P. (2009). Understanding PREMIS. *D-Lib Magazine*, *15*(3-4).

Caron, B., Ledoux, T., Reecht, T. et Tramoni, J.-P. (2015, novembre). *Experiment, document and decide : a collaborative approach to preservation planning at the BnF*. Communication présentée à iPRES 2015 12th International Conference on Digital Preservation, Chapell Hill, États-Unis. Repéré à https://hal-bnf.archives-ouvertes.fr/hal-01288699/document

Copeland, B. J. (2016). Artificial intelligence (AI). Dans *Encyclopædia Britannica.* Repéré à https://www.britannica.com/technology/artificial-intelligence

Coyle, K. (2010). Library data in a modern context. *Library Technology Reports, 46*(1), 5-13.

Creative Commons. (s.d.). Creative Commons – Attribution 4.0 international – CC BY 4.0. Repéré à https://creativecommons.org/licenses/by/4.0/deed.fr

DCMI. (2012). Dublin Core metadata element set, version 1.1. Repéré à http://dublincore.org/documents/dces/

Denton, W. (2007). FRBR and the history of cataloguing. Dans A.G. Taylor (dir.), *Understanding FRBR : What it is and how it will affect our retrieval* (p. 35-57). Westport, Connecticut : Libraries Unlimited.

DeWeese, P. K. et Segal, D. (2015). *Libraries and the semantic web*. San Rafael, California : Morgan & Claypool.

Etalab. (s.d.). La mission Etalab | le blog de la mission Etalab. Repéré à https://www.etalab.gouv.fr/qui-sommes-nous

Farkas, M. G. (2007). *Social software in libraries : Building collaboration, communication, and community online.* Information Today Inc.

Feigenbaum, L. et Prud'hommeaux, E. (2013). SPARQL by example - Cambridge Semantics. Repéré à http://www.cambridgesemantics.com/semantic-university/sparql-by-example

Foster, A. et Ford, N. (2003). Serendipity and information seeking : An empirical study. *Journal of Documentation*, *59*(3), 321-340. doi : 0.1108/00220410310472518

Gangemi, A. et Presutti, V. (2006, mai). *The bourne identity of a web resource.* Communication présentée à Workshop Identity, Reference and the Web (IRW) à la conference WWW, Écosse. Repéré à http://wwwconference.org/proceedings/www2006/www.ibiblio.org/hhalpin/irw2006/vpresutti.pdf

Gartner. (2016). Internet of things. Repéré à http://www.gartner.com/it-glossary/internet-of-things/

Gonzalez, B.M. (2014). Linking libraries to the web : Linked data and the future of the bibliographic record. *Information Technology and Libraries*, *33*(4), 10-22. doi: 10.6017/ital.v33i4.5631

Google. (s.d.) JSON-LD | Google Schemas | Google Developers. Repéré à https://developers.google.com/schemas/formats/json-ld

Google. (2010). Official Google blog : deeper understanding with metaweb. Repéré à https://googleblog.blogspot.ca/2010/07/deeper-understanding-with-metaweb.html

Gouvernement du Québec. (2016-a). À propos : Plan culturel numérique. Repéré à http://culturenumerique.mcc.gouv.qc.ca/a-propos/

Gouvernement du Québec. (2016-b). 06 – Aider le réseau de la culture à s'approprier les technologies du Web sémantique afin de maximiser la présence des données culturelles québécoises dans le Web. Repéré à http://culturenumerique.mcc.gouv.qc.ca/aider-le-reseau-de-la-culture-a-sapproprier-les-technologies-du-web-semantique-afin-de-maximiser-la-presence-des-donnees-culturelles-quebecoises-dans-le-web-banq/

Gruber, T. (s.d.). Ontology (computer science) – definition in encyclopedia of database systems. Repéré à http://tomgruber.org/writing/ontology-definition-2007.htm

Guha, R. V. (2011). Introducing schema.org : Search engines come together for a richer web. Repéré à https://googleblog.blogspot.ca/2011/06/introducing-schemaorg-search-engines.html

Guha, R.V., Brickley, D. et Macbeth, S. (2016). Schema.org : Evolution of structured data on the web. *Communications of the ACM*, *59*(2), 44-51. doi: 10.1145/2844544

Guinard, D., Trifa, V., Mattern, F. et Wilde, E. (2011). From the internet of things to the web of things : Resource-oriented architecture and best practices. Dans D. Uckelmann, M. Harrison et F. Michahelles (dir.), *Architecting the Internet of Things* (p. 97-129). Springer Berlin Heidelberg.

Hannemann, J. et Kett, J. (2010, Août). *Linked data for libraries*. Communication présentée au World Library and Information Congress: 76th IFLA General Conference and Assembly, Gothenburg, Suède. Repéré à http://conference.ifla.org/past-wlic/2010/149-hannemann-en.pdf

Hart, G. et Dolbear, C. (2013). *Linked data : a geographic perspective*. Boca Raton : CRC Press.

Hawkins, L. (2015). The semantic web and the BIBFRAME initiative. *Serials Review*, *41*(2), 106-107. doi: 10.1080/00987913.2015.1030962

Heath, T. (2009, 2 mars). Linked data? Web of data? Semantic web? WTF? [Billet de blogue]. Repéré à http://tomheath.com/blog/2009/03/linked-data-web-of-data-semantic-web-wtf/

Hendler, J. (2009, 16 juin). What is the semantic web really all about? [Billet de blogue]. Repéré à http://www.scilogs.com/web_science/what-is-the-semantic-web-really-all-about/

Hicks, D. (2013). *Technology and professional identity of librarians : The making of the cybrarian*. Hershey, États-Unis : IGI Global.

Hyvönen, E. (2012). *Publishing and using cultural heritage linked data on the semantic web*. San Rafael, California : Morgan & Claypool.

Howarth, L. C. (2012). FRBR and linked data : Connecting FRBR and linked data. *Cataloging & Classification Quarterly*, *50*(5-7), 763-776.

IBM. (2003-a). IBM completes acquisition of rational software. Repéré à http://www-03.ibm.com/press/us/en/pressrelease/314.wss

IBM. (2003-b). Risk reduction with the RUP phase plan. Repéré à http://www.ibm.com/developerworks/rational/library/1826.html#N100E4

IFLA/UNESCO. (1994). Manifeste de l'IFLA/UNESCO sur la bibliothèque publique 1994. Repéré à http://www.ifla.org/FR/publications/manifeste-de-liflaunesco-sur-la-biblioth-que-publique-1994

IFLA Study Group on the Functional Requirements for Bibliographic Records. (1998). *Functional Requirements for Bibliographic Records*. Munich : Saur.

Institut national de la statistique et des études économiques. (2016). Population totale par sexe et âge au 1er janvier 2016, France. Repéré à http://www.insee.fr/fr/themes/detail.asp?reg_id=0&ref_id=bilan-demo&page=donnees-detaillees/bilan-demo/pop_age2b.htm

Isaac, A. (2012). Les référentiels : typologie et interopérabilité. Dans Calderan, L., Laurent, P., Lowinger, H. et Millet, J. (dir.), *Le document numérique à l'heure du Web de données : séminaire INRIA, Carnac, 1er-5 octobre 2012* (p.105-134). Paris : ADBS Éditions.

ITU. (2016). Internet of things global standards initiative. Repéré à http://www.itu.int/en/ITU-T/gsi/iot/Pages/default.aspx

Jacquesson, A. (1992). *L'informatisation des bibliothèques : histoire, stratégie et perspectives*. Paris: Éditions du Cercle de la Librarie.

Koontz, C. et Mon, L. (2014). *Marketing and social media : A guide for libraries, archives and museums*. Lanham : Rowman & Littlefield.

Kopetz, H. (2011). Internet of things. Dans *Real-time systems* (p.307-323). Springer US.

Kruchten, P. (2000). *Introduction au rational unified process*. Paris: Eyrolles.

Le Bœuf, P. (2013, août). *Customized OPACs on the semantic web : the OpenCat prototype*. Communication présentée au 78th IFLA General Conference and Assembly, Singapour. Repéré à http://files.dnb.de/svensson/UILLD2013/UILLD-submission-3-formatted-final.pdf

Léopold, D. (1907). *Recherches sur la librairie de Charles V – Vol. 1.* Paris : H. Champion.

Leroux, E. et al. (2009). Les professions et les institutions. Dans J. M. Salaün et C. Arsenault (dir.), *Introduction aux sciences de l'information* (p. 16-52). Montréal, Québec : Presses de l'Université de Montréal.

Library of Congress. (2008). MARC to Dublin Core crosswalk. Repéré à https://www.loc.gov/marc/marc2dc.html

Library of Congress. (2012-a). The Library of Congress announces modeling initiative (May 22, 2012). Repéré à http://www.loc.gov/bibframe/news/bibframe-052212.html

Library of Congress. (2012-b). Bibliographic Framework as a web of data : linked data model and supporting services. Repéré à http://www.loc.gov/bibframe/pdf/marcld-report-11-21-2012.pdf

Library of Congress. (2016). Overview of the BIBFRAME 2.0 model. Repéré à http://www.loc.gov/bibframe/docs/bibframe2-model.html

Liew, C.L., Wellington, S., Oliver G. et Perkins, R. (2015). Social media in libraries and archives : Applied with caution. *Canadian Journal of Information and Library Science*, *39*(3/4), 377-396.

Lytras, M. D. et Sicilia, M. A. (2005). The knowledge society : A manifesto for knowledge and learning. *International Journal of Knowledge and Learning*, *1*(1-2), 1-11.

Merton, R.K. et Barber, E. (2004). *The travels and adventures of serendipity : A study in sociological semantics and the sociology of science*. Princeton University Press.

Mikhalenko, P. (2005). Introducing SKOS. Repéré à http://www.xml.com/pub/a/2005/06/22/skos.html

Ministère du Travail, de l'Emploi, de la Formation professionnelle et du Dialogue social. (2013). Emplois d'avenir. Repéré à http://travail-emploi.gouv.fr/emplois-d-avenir,2189/presentation,2259/emplois-d-avenir,2189/rubrique-technique,2193/presentation,2278/presentation,16015.html

Munassar, N. M. A. et Govardhan, A. (2010). A comparison between five models of software engineering. *International Journal of Computer Science Issues*, *5*, 95-101.

Noy, N. F. et McGuinness, D. L. (2001). Ontology development 101 : A guide to creating your first ontology. Repéré à http://protege.stanford.edu/publications/ontology_development/ontology101.pdf

O'Reilly, T. (2005). What is web 2.0 : design patterns and business models for the next generation of software. Repéré à http://www.oreilly.com/pub/a/web2/archive/what-is-web-20.html

Powell, W. W. et Snellman, K. (2004). The knowledge economy. *Annual review of sociology*, *30*, 199-220. doi: 10.1146/annurev.soc.29.010202.100037

Rahm, E. et Do, H. H. (2000). Data cleaning : Problems and current approaches. *IEEE Data Engineering Bulletin*, *23*(4), 3-13.

Rao, K. N. et Babu, K. H. (2001). Role of librarian in internet and world wide web environment. *Information Science*, *4*(1), 25-34.

Rational Software. (1998). Rational unified process. Best practices for software development teams. Repéré à https://www.ibm.com/developerworks/rational/library/content/03July/1000/1251/1251_bestbestpract_TP026B.pdf

RFN. (s.d.-a). Mission. Repéré à http://www.rfnum.org/pages/a_propos/mission.html

RFN. (s.d.-b). Données ouvertes du RFN : licence d'utilisation. Repéré à http://www.rfnum.org/pages/donnees_ouvertes/licence.html

Sami, M. (2012, 21 mars). Choosing the right software development life cycle model [Billet de blogue]. Repéré à https://melsatar.wordpress.com/2012/03/21/choosing-the-right-software-development-life-cycle-model/

Schema.org. (s.d.). About Schema.org. Repéré à http://www.schema.org/docs/faq.html

Schilling, V. (2012). Transforming library metadata into linked library data. Repéré à http://www.ala.org/alcts/resources/org/cat/research/linked-data

Schillinger, R. (2011). *Semantic service oriented architectures in research and practice*. Cologne : Josef Eul Verlag.

Schmachtenberg, M., Bizer, C., Jentzsch, A. et Cyganiak, R. (2014). The linking open data cloud diagram. Repéré à http://lod-cloud.net/

Sepetjean, S. et Graff, E. (2011). Le dépôt légal en France. *Les cahiers de propriété intellectuelle*, *1*(23), 169-186.

Shadbolt, N., Hall, W. et Berners-Lee, T. (2006). The semantic web revisited. *Intelligent Systems, IEEE*, *21*(3), 96-101. doi: 10.1109/MIS.2006.62

Simon, A., Di Mascio, A., Michel V. et Peyrard, S. (2014, août). *We grew up together : data.bnf.fr from the BnF and Logilab perspectives*. Communication présentée au IFLA World Library & Information Congress, Lyon, France. Repéré à http://ifla2014-satdata.bnf.fr/pdf/iflalld2014_submission_Simon_DiMascio_Michel_Peyrard.pdf

Simon, A., Wenz, R., Michel, V. et Di Mascio, A. (2013, mai). *Publishing bibliographic records on the web of data : Opportunities for the BnF (French National Library)*. Communication présentée au 10th International Conference, ESWC 2013, Montepellier, France. Repéré à http://eswc-conferences.org/sites/default/files/papers2013/simon.pdf

Synak, M., Dabrowski, M. et Kruk, S.R. (2009). Semantic web and ontologies. Dans Kruk, S.R. et McDaniel, B. (dir.), *Semantic digital libraries* (p. 41-54). Berlin : Springer.

Stanford University Libraries. (s.d.). Stanford prize for innovation in research libraries (SPIRL). Repéré à http://library.stanford.edu/projects/stanford-prize-innovation-research-libraries-spirl

Statistique Canada. (2015). Population par année, par province et territoire (nombre). Repéré à http://www.statcan.gc.ca/tables-tableaux/sum-som/l02/cst01/demo02a-fra.htm

Stephens, M. et Collins, M. (2007). Web 2.0, library 2.0, and the hyperlinked library. *Serials Review*, *33*(4), 253-256.

Stuart, D. (2011). *Facilitating access to the web of data : A guide for librarians*. London : Facet.

Swanson, T.A. (2012). *Managing social media in libraries : Finding collaboration, coordination and focus.* Oxford : Chandos Pub.

Thomsett-Scott, B.C. (2014). *Marketing with social media : a LITA guide*. Chicago : ALATechsource, an imprint of the American Library Association.

Tillett, B. (2013). RDA and the semantic web, linked data environment. *JLIS.it*, *4*(1), 139-145. doi:10.4403/jlis.it-6303

Turner, J., Dufour, C., Laplante, A., Leroux, E. et Salaün, J. M. (2009). Les pratiques des utilisateurs. Dans J. M. Salaün et C. Arsenault (dir.), *Introduction aux sciences de l'information* (p. 159-182). Montréal, Québec : Presses de l'Université de Montréal.

UML. (2005). What is UML | Unified modeling language. Repéré à http://www.uml.org/what-is-uml.htm

UNESCO. (2005). *Vers les sociétés du savoir*. Paris, France : Éditions UNESCO.

Van Hooland, S. et Verborgh, R. (2014). *Linked data for libraries, archives and museums : How to clean, link and publish your metadata*. Chicago : Neal-Schuman, an imprint of the American Library Association.

Vatant, B. (2008). Des métadonnées à la description de ressources : les langages du web sémantique. Dans Calderan, L., Hidoine, B. et Millet, J. (dir.), *Métadonnées : mutations et perspectives : séminaire INRIA, 29 septembre-3 octobre 2008-Dijon* (p. 163-194). Paris : ADBS Éditions.

VIAF. (s.d.). VIAF : Fichier d'autorité international virtuel. Repéré à http://viaf.org/

Villazón-Terrazas, B., Vilches-Blázquez, L. M., Corcho, O. et Gómez-Pérez, A. (2011). Methodological guidelines for publishing government linked data. Dans D. Wood (dir.), *Linking government data* (p.27-49). New York : Springer.

W3C. (2007). Latest layercake diagram (PNG). Repéré à https://www.w3.org/2007/03/layerCake.png

W3C. (2008). Cool URIs for the semantic web. Repéré à https://www.w3.org/TR/cooluris/

W3C. (2013). W3C data activity. Repéré à http://www.w3.org/2013/data/

W3C. (2014-a). RDF 1.1 Turtle. Repéré à https://www.w3.org/TR/turtle/

W3C. (2014-b). Best practices for publishing linked data. Repéré à https://www.w3.org/TR/ld-bp/

W3C. (2015). Vocabularies. Repéré à http://www.w3.org/standards/semanticweb/ontology

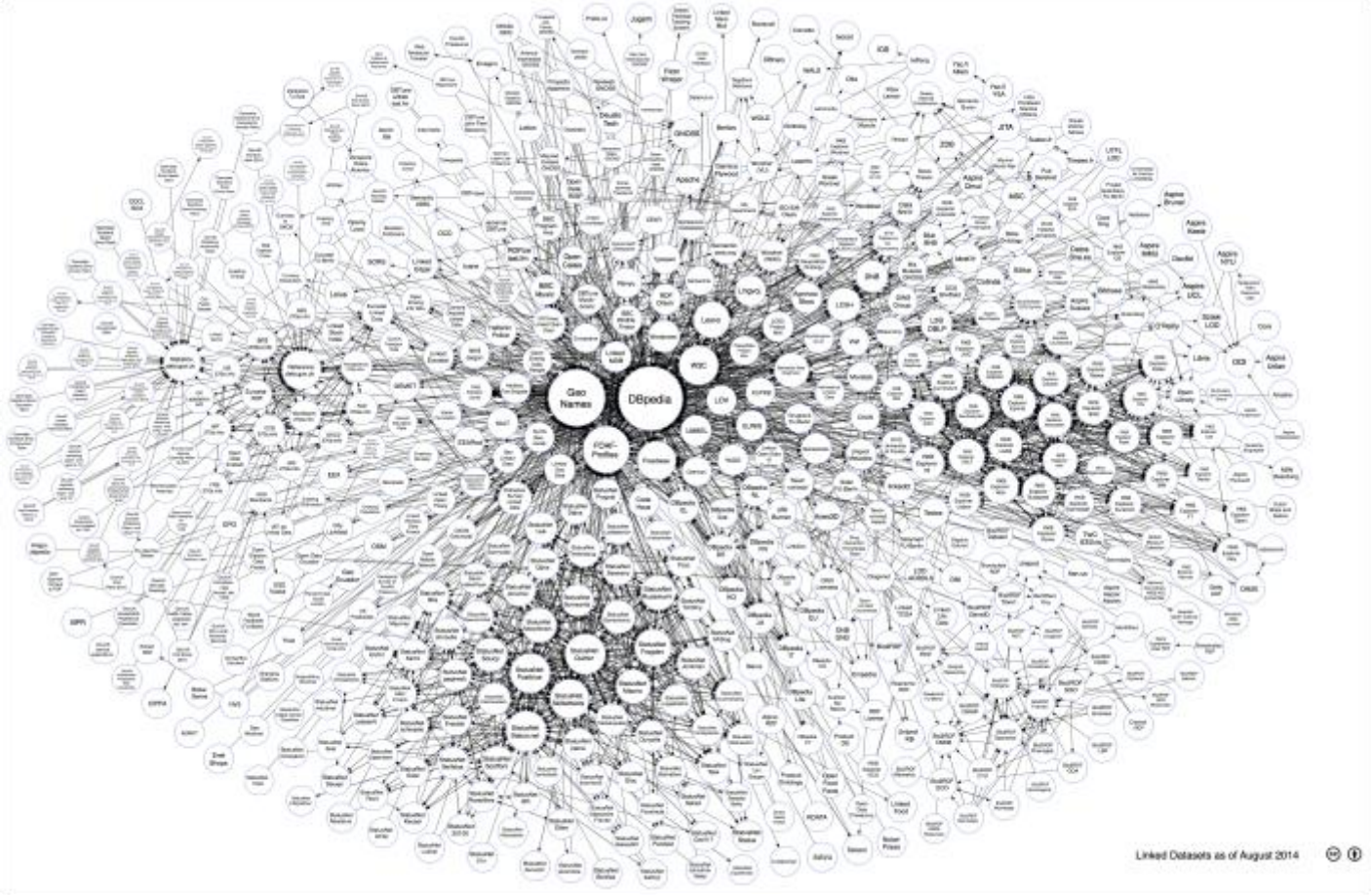
Wikidata. (2015). Wikidata : introduction. Repéré à https://www.wikidata.org/wiki/Wikidata:Introduction/fr

Willer, M. et Dunsire, G. (2013). *Bibliographic information organization in the semantic web.* Witney, Oxford : Chandos Publishing.

Zaino, J. (2013). Semantic web gets closer to the internet of things. Repéré à http://www.dataversity.net/34702/

Zengenene, D., Casarosa, V. et Meghini, C. (2014). Towards a methodology for publishing library linked data. Dans T. Catarci, N. Ferro et A. Poggi (dir.), *Bridging between cultural heritage institutions* (p. 81-92). Berlin : Springer Berlin Heidelberg.

Annexe 1 – ***Linking Open Data cloud diagram***



Source : Linking Open Data cloud diagram 2014, par Max Schmachtenberg, Christian Bizer, Anja Jentzsch et Richard Cyganiak. http://lod-cloud.net/

Annexe 2 – Exemple simple de triplets RDF en tableau et en graphe

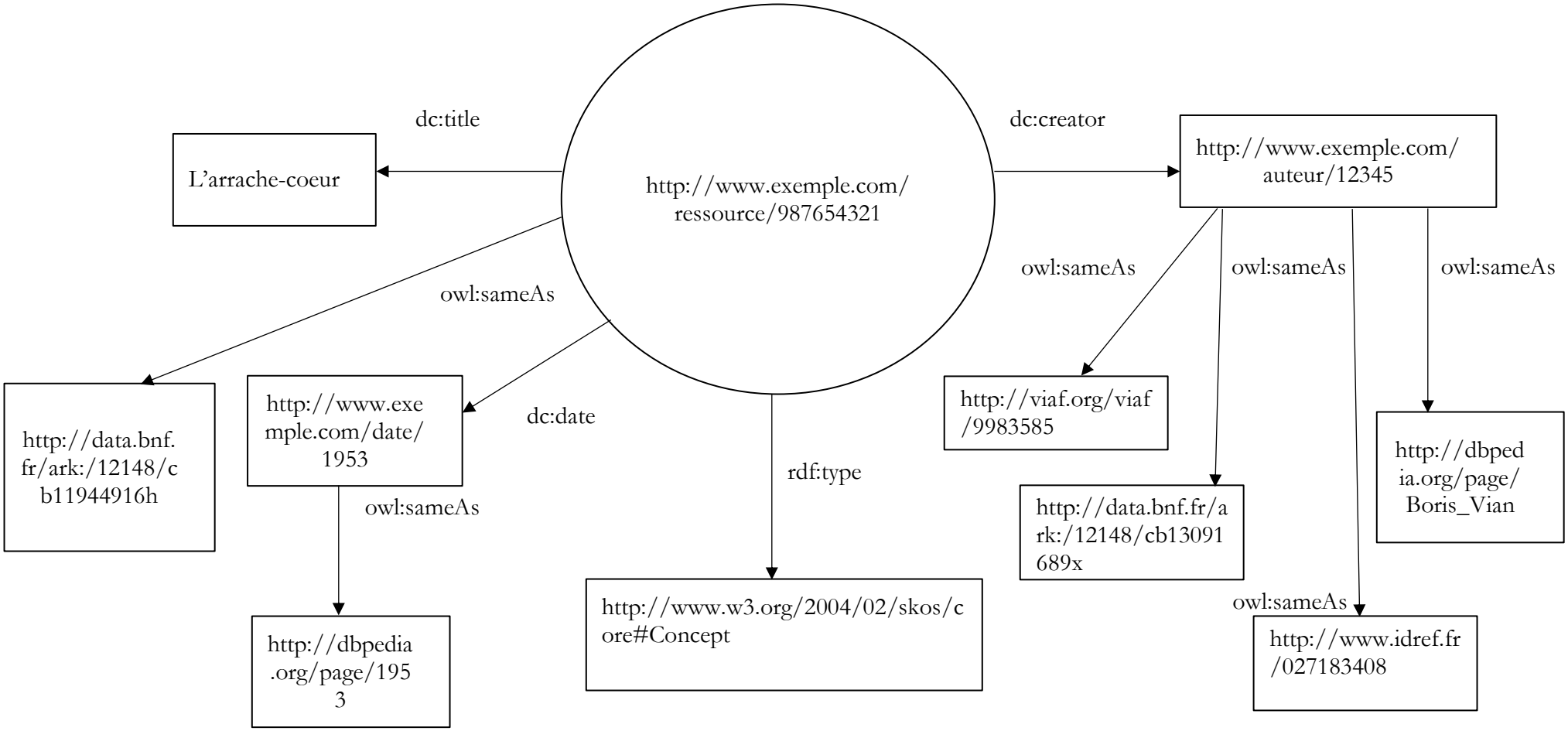
Représentations tabulaires :

http://www.exemple.com/ressource/987654321 étant l'identifiant pour la ressource *L'arrache-cœur* de Boris Vian.
http://www.exemple.com/auteur/12345 étant l'identifiant pour la ressource Boris Vian.

Sujet	Prédicat	Objet (interne)
http://www.exemple.com/ressource/987654321	rdf:type	http://www.w3.org/2004/02/skos/core#Concept
http://www.exemple.com/ressource/987654321	dc:title	L'arrache-coeur
http://www.exemple.com/ressource/987654321	dc:creator	http://www.exemple.com/auteur/12345
http://www.exemple.com/ressource/987654321	dc:date	http://www.exemple.com/date/1953

Sujet	Prédicat	Objet
http://www.exemple.com/ressource/987654321	owl:sameAs	http://data.bnf.fr/ark:/12148/cb11944916h
http://www.exemple.com/auteur/12345	owl:sameAs	http://viaf.org/viaf/9983585 http://data.bnf.fr/ark:/12148/cb13091689x http://www.idref.fr/027183408 http://dbpedia.org/page/Boris_Vian
1953	owl:sameAs	http://dbpedia.org/page/1953

Représentation en graphe :



Annexe 3 – Glossaire

Agent logiciel : Logiciel agissant de manière autonome qui effectue des recherches, établit des liens entre les différents résultats obtenus et les compare pour présenter le résultat le plus pertinent.

Alignement : Mise en correspondance des concepts afin de créer des liens entre les ressources internes et externes. Il s'agit ainsi de liens entre des entités qui sont équivalentes, similaires ou connexes au niveau du sens. Ces liens se font à l'aide de vocabulaires, d'ontologies, de schémas ou de jeux de données externes.

Application Programming Interface **(API) :** Interface de programmation qui permet d'accéder à l'ensemble de règles qui composent une application. Un développeur peut ainsi y accéder et procéder à l'échange de données.

Classe : La classe permet d'obtenir de l'information sur la nature d'une ressource.

Données structurées internes : Contenu sémantique ajouté à même les pages HTML afin d'en faciliter le repérage par les moteurs de recherche. Il s'agit d'une application du Web sémantique.

Dump **:** Lien cliquable permettant le téléchargement souvent massif de jeux de données.

Espace de nom : Lieu virtuel où l'on retrouve l'organisation et la définition de termes particuliers.

Instance : Ressource faisant partie d'une classe donnée.

Internationalized Resource Identifier **(IRI) :** Identifiant permettant l'utilisation de tous les alphabets utilisés dans le monde.

Licence d'utilisation : Contrat dans lequel sont stipulées les conditions d'utilisation, de modification et de diffusion des jeux de données.

Métadonnée : Donnée présentant de l'information à propos d'une autre donnée.

Modèle de données : Modèle visant à définir la structure et la sémantique des données contenues dans un système.

Négociation de contenu : Mécanisme qui permet de proposer pour une même ressource différentes visualisations en fonction de la provenance de la requête.

Ontologie : Ensemble structuré visant à définir les concepts et les relations afin de classifier l'utilisation de certains termes dans une application donnée, caractériser les relations entre les ressources et clarifier les contraintes quant à l'utilisation d'un terme.

Point d'accès SPARQL : Service qui respecte le protocole SPARQL et qui présente la possibilité d'effectuer des requêtes dans une base de données.

Propriété : La propriété donne de l'information sur le prédicat du triplet, donc sur la relation qui unit deux ressources ou une ressource et un littéral.

***Resource Description Framework* (RDF) :** Langage de base du Web sémantique ayant comme objectif de décrire les ressources et leurs métadonnées. Un document RDF est basé sur la formation de triplets.

***Resource Description Framework Schema* (RDFS) :** Langage de base de représentation de connaissances qui fournit des éléments pour la construction d'ontologies et de vocabulaires pour la description de ressources.

***Semantic Web Stack* :** Illustration permettant la visualisation de la hiérarchisation des technologies et standards qui construisent le Web sémantique

Schéma de métadonnées : Schéma visant à définir un ensemble fini et formellement défini de métadonnées utilisées pour décrire des ressources.

***Simple Knowledge Organization System* (SKOS) :** Modèle de représentation de thésaurus, de classifications ou d'autres vocabulaires contrôlés. SKOS permet aussi de les mettre en correspondance.

***SPARQL Protocol and RDF Query Language* (SPARQL) :** Langage de requête permettant d'effectuer des recherches au sein de données encodées en RDF, de les récupérer et de les manipuler.

Syntaxe de sérialisation : Syntaxe permettant l'encodage de données et le partage de celles-ci entre les systèmes.

***Triplestore* :** Base de données conçue pour stocker et récupérer des données en RDF. On retrouve donc dans un *triplestore* uniquement des triplets.

Triplet RDF : Énoncé créé à partir d'un sujet (ressource décrite), d'un prédicat (propriété de la ressource ou relation) et d'un objet (donnée ou autre ressource).

***Uniform Resource Identifier* (URI) :** Suite alphanumérique qui identifie de manière univoque une ressource physique ou abstraite.

***Uniform Resource Name* (URN) :** URI permettant d'identifier une ressource durant toute la durée de son existence, et ce, indépendamment de sa location ou de son accessibilité.

URI déréférençable : URI qui se définit principalement par le fait qu'il met en pratique le mécanisme de négociation de contenu.

Vocabulaire : Schéma permettant l'organisation des connaissances de manière standardisée afin de faciliter la recherche d'information.

Web 2.0 : Mouvement découlant du *World Wide Web* misant sur la participation des usagers, la création de contenu et l'interopérabilité.

Web de données : Application du Web sémantique visant la publication de données en format RDF afin de permettre la création de liens entre celles-ci et leur compréhension par la machine.

***Web Ontology Language* (OWL) :** Langage de représentation des connaissances qui permet de définir et décrire des ontologies.

Web sémantique : Extension du Web tel qu'on le connaît basée sur des standards définis par le W3C et qui vise la publication de données dans un format standardisé afin de les exploiter de manière plus efficace.

***World Wide Web* :** Système basé sur des standards tels que l'hypertexte et qui permet la recherche et la visualisation d'informations via Internet.

***World Wide Web Consortium* (W3C) :** Organisme de standardisation pour le *World Wide Web*.