h e g

Haute école de gestion
Genève

Modélisation et transformation des métadonnées de RERO en Linked Open Data

réro

Réseau des bibliothèques de Suisse occidentale
Westschweizer Bibliotheksverbund
Rete delle biblioteche della Svizzera occidentale
Library Network of Western Switzerland

Travail de Master réalisé en vue de l'obtention du Master HES

par :

Nicolas PRONGUÉ

Directeur du travail de Master :

Dr René SCHNEIDER, professeur HES

Genève, 29 août 2014

Haute école de gestion de Genève (HEG-GE)

Filière Information documentaire

Hes·so GENÈVE
Haute Ecole Spécialisée
de Suisse occidentale

Déclaration

Ce travail de Master est réalisé dans le cadre de l'examen final de la Haute école de gestion de Genève, en vue de l'obtention du titre de Master of Science en information documentaire.

L'étudiant a envoyé ce document par email à l'adresse remise par son directeur de travail de Master pour analyse par le logiciel de détection de plagiat URKUND, selon la procédure détaillée à l'URL suivante : http://www.urkund.com/fr/student

L'étudiant accepte, le cas échéant, la clause de confidentialité. L'utilisation des conclusions et recommandations formulées dans le travail de Master, sans préjuger de leur valeur, n'engage ni la responsabilité de l'auteur, ni celle du directeur du travail de Master, du juré ou de la HEG.

« J'atteste avoir réalisé seul le présent travail, sans avoir utilisé des sources autres que celles citées dans la bibliographie. »

Fait à Genève, le 29 août 2014

Nicolas Prongué

Remerciements

Merci à Miguel Moreira, représentant de RERO dans ce projet, pour son enthousiasme et son professionnalisme, et la collaboration efficace qui en a découlé.

Merci à René Schneider, directeur de ce travail de Master, pour sa disponibilité et ses conseils pédagogiques pertinents.

Merci à mes parents pour la relecture attentive de ce rapport.

Résumé

A l'arrivée du web, une multitude de documents et d'informations du monde entier se sont peu à peu reliés et interconnectés. Un réseau mondial de connaissances est né, basé sur des standards internationaux et intercommunautaires. Toutefois, les données des bibliothèques, à cause de leur format d'enregistrement particulier, n'ont pas pleinement intégré ce réseau, et des portails de recherche spécialisés ont été développés pour y accéder. A présent, de nouveaux standards techniques, permettant de meilleurs liens et une interopérabilité augmentée, ont été créés ; le web sémantique n'est plus une utopie, il est aux portes des bibliothèques.

Ce travail a pour but de préparer les données de RERO, disponibles en format MARC21, à la publication en Linked Open Data, afin de les intégrer au web sémantique. Cela permettra entre autres de les rendre plus visibles et plus facilement réutilisables, et à terme de développer de nouvelles fonctionnalités pour les utilisateurs. Ce processus se décompose en plusieurs étapes : revue de la littérature, analyse des données de base, modélisation, mapping et ajout de liens externes. La modélisation inclut la sélection d'un modèle de base, ainsi que l'attribution d'informations de provenance et d'identifiants pérennes aux données. Le mapping consiste quant à lui à choisir des ontologies pertinentes, établir des correspondances entre le format de départ et le format de destination, et formuler des règles de conversion. L'ajout de liens externes implique le choix de référentiels pertinents du web sémantique, en fonction des possibilités de liens offertes par la structure et le niveau de contrôle des données de RERO.

Au final, le modèle développé se décline en six tables de mapping – une pour chaque type d'entité identifié dans les données du réseau – et environ 130 règles de conversion. Il se base sur une douzaine d'ontologies, toutes déjà existantes.

Ce projet a mis en évidence certains aspects problématiques de l'encodage traditionnel des données bibliographiques, relatifs au format d'enregistrement et aux règles de catalogage utilisés. Le passage à de nouveaux standards de description, un processus en cours dans le domaine des bibliothèques, vise entre autres la résolution de ces difficultés.

Mots-clés : bibliothèques, conversion, données, Linked Open Data, mapping, métadonnées, modélisation, web sémantique

Table des matières

Liste des tableaux

Liste des figures

Sigles et acronymes

AACR2	Anglo-American Cataloguing Rules, 2e édition
API	Application Programming Interface (Interface de programmation des applications)
BIBFRAME	Bibliographic Framework
BIBO	Bibliographic Ontology
BNB	British National Bibliography (Bibliographie nationale britannique)
BNE	Biblioteca Nacional de España (Bibliothèque nationale d'Espagne)
BnF	Bibliothèque nationale de France
CDU	Classification décimale universelle
CURIE	Compact URI
DC	Dublin Core
DNB	Deutsche Nationalbibliothek (Bibliothèque nationale allemande)
EDM	Europeana Data Model
FOAF	Friend Of A Friend
FRAD	Functional Requirements for Authority Data (Fonctionnalités requises des données d'autorité)
FRBR	Functional Requirements for Bibliographic Records (Fonctionnalités requises des notices bibliographiques)
FRSAD	Functional Requirements for Subject Authority Data (Fonctionnalités requises des données d'autorité matière)
GND	Gemeinsame Normdatei (fichier d'autorités commun)
HBZ	Hochschulbibliothekszentrum (Centre des bibliothèques des hautes écoles de Rhénanie-du-Nord-Westphalie)
HTML	HyperText Markup Language
HTTP	HyperText Transfer Protocol

IFLA	International Federation of Library Associations (Fédération internationale des associations et institutions de bibliothèques)
IRI	Internationalized Resource Identifier
ISBN	International Standard Book Number
JSON	JavaScript Object Notation
JSON-LD	JavaScript Object Notation for Linked Data
LCSH	Library of Congress Subject Headings
LOC	Library of Congress (Bibliothèque du Congrès, USA)
LOD	Linked Open Data
MARC	MAchine Readable Cataloging
OCLC	Online Computer Library Center
OWL	Web Ontology Language
RAMEAU	Répertoire d'autorité-matière encyclopédique et alphabétique unifié
RDA	Resource Description and Access
RDF	Resource Description Framework
RDFa	RDF in attributes
RDFS	RDF Schema
RDF/XML	Resource Description Framework / eXtensible Markup Language
RERO	Réseau romand : Réseau des bibliothèques de Suisse occidentale
SKOS	Simple Knowledge Organization System
SPARQL	SPARQL Protocol and RDF Query Language
SUDOC	Système universitaire de documentation (catalogue des institutions de l'enseignement supérieur et de la recherche françaises)
TTL	Turtle, ou Terse RDF Triple Language
URI	Uniform Resource Identifier
VIAF	Virtual International Authority File

VoID	Vocabulary of Interlinked Datasets
W3C	World Wide Web Consortium
VRA	Visual Resources Association
XML	eXtensible Markup Language

1. Introduction

L'ère du numérique et l'essor du web plongent les bibliothèques dans une phase de profonde remise en question sur de nombreux plans. S'agissant de leurs données bibliographiques, le format traditionnel d'enregistrement, nommé MARC, peine à s'adapter à l'environnement web en rapide évolution. Les supports d'information numériques rendent peu à peu les règles de catalogage en vigueur obsolètes. Parallèlement, de nouvelles tendances voient le jour. Les standards naissants du web sémantique promettent flexibilité et interopérabilité dans la gestion des données. Le mouvement de l'Open Data convainc de plus en plus d'institutions publiques à ouvrir leurs données, à des fins de transparence et de partage.

Dans ce contexte, les bibliothèques cherchent activement à faire évoluer leur modèle actuel de gestion des métadonnées. Une telle opération représente un défi de poids, car elle implique le développement et l'instauration dans le monde entier de nouveaux standards, notamment concernant les règles de catalogage, le modèle de données et le format d'enregistrement. Les bibliothèques souhaitent ainsi prendre part aux évolutions de l'ère numérique et s'intégrer pleinement au web des données. Les Linked Open Data (LOD), des données ouvertes et liées enregistrées dans un format du web sémantique, sont une clé dans ce changement de paradigme.

S'insérant dans cette mouvance, RERO, le Réseau des bibliothèques de Suisse occidentale, a pris la décision de publier ses métadonnées en LOD. En vue de ce projet, ce travail de Master a pour but de préparer les données en les modélisant, en établissant des correspondances (mapping) entre le format de départ et le format de destination et en formulant des règles de conversion. Il se concentre sur ces aspects d'un point de vue bibliothéconomique et théorique, sans aborder en détail la mise en œuvre technique que cela implique.

Dans la première partie de ce rapport, le web sémantique, ainsi que les principaux concepts sur lesquels il se base, sont décrits en général. De plus, un regard est porté sur les métadonnées en bibliothèque et les défis auxquels elles font face actuellement. La deuxième partie introduit brièvement RERO, l'institution responsable du projet de publication. Ensuite sont présentés la méthodologie de travail adoptée, le processus détaillé tel qu'il s'est déroulé ainsi que les réalisations effectuées dans ce travail. Enfin ce rapport discute les résultats obtenus, l'impact que le projet peut avoir sur RERO et les perspectives qu'il lui offre.

2. Linked Open Data et bibliothèques

2.1 Linked Open Data

2.1.1 Définitions

Les Linked Open Data (LOD) sont des données ouvertes et liées, publiées dans le respect des normes du web sémantique. Le terme Linked Open Data résulte de la fusion de deux concepts distincts : Linked Data et Open Data.

Premièrement, Linked Data est un terme créé en 2006 par Tim Berners-Lee dans le contexte du web sémantique.

Le web sémantique est défini par Emmanuelle Bermès, spécialiste française dans le domaine, comme « un ensemble de technologies et de normes, développées par l'organisme de normalisation du web, le W3C, afin de faciliter le traitement des données par des machines » (BnF 2014a). Il se veut une extension du web actuel, dit w*eb of documents*, vers un web centré sur les données : « The semantic web is a web of data » (W3C 2013a). Cette vision du web est également issue d'une idée de Berners-Lee, qui l'a présentée pour la première fois en 2001. Il voulait créer un web qui permette aux ordinateurs et aux logiciels de travailler en coopération avec les personnes et de développer la connaissance humaine, au moyen de l'interprétation automatique de données structurées (Berners-Lee, Hendler, Lassila 2001). Trop complexe et utopique, l'initiative n'a pas tout de suite obtenu le succès escompté. Le phénomène prit réellement de l'ampleur en 2006, lorsque le célèbre informaticien apposa une nouvelle étiquette au web sémantique, les Linked Data, dans le but de populariser le concept en le présentant de manière plus pragmatique (Pohl, Danowski 2013, p. 4-5).

Créer des Linked Data consiste à utiliser les standards du web sémantique pour établir, entre plusieurs jeux de données, des liens *définis*, c'est-à-dire des liens ayant une signification. Par exemple, un lien défini décrira l'entité *Berne* comme étant la capitale de l'entité *Suisse* d'une manière reconnue universellement, y compris par des machines, ce qui lui donne un potentiel énorme. A l'inverse, dans le web traditionnel, la nature d'un lien hypertexte entre deux documents est à déduire par le lecteur humain. Basées sur le standard RDF, les Linked Data ont pour but de fournir des données interopérables pouvant être traitées automatiquement par des ordinateurs (W3C 2009, chap. 1.1; Bizer, Heath, Berners-Lee 2009, chap. 2).

Le second concept est l'Open Data, ou données ouvertes en français. Il est défini ainsi par l'Open Knowledge Foundation (2012) :

> *« Une donnée ouverte est une donnée qui peut être librement utilisée, réutilisée et redistribuée par quiconque - sujette seulement, au plus, à une exigence d'attribution et de partage à l'identique. »*

Par déduction, cette définition stipule donc qu'une donnée ouverte doit pouvoir être réutilisée à des fins commerciales également. Pour faciliter la réutilisation des données, trois aspects doivent être pris en compte (Chignard 2012, p. 13-14) :

1. Aspect technique : un format le plus ouvert possible
2. Aspect juridique : une licence ouverte
3. Aspect économique : des redevances nulles ou limitées

L'Open Data s'inscrit dans une mouvance plus générale d'ouverture des connaissances, incluant entre autres l'Open Source (pour les logiciels), l'Open Innovation (pour la recherche) ou encore l'Open Access (pour les ressources d'information). Le concept est par ailleurs étroitement lié à celui de l'Open Government Data, qui concerne les données produites par des institutions de droit public.

L'interopérabilité technique et la puissance sémantique des Linked Data, couplées à la libre utilisation juridique et la gratuité de l'Open Data, forment la puissance du concept de Linked Open Data. L'aspect ouvert des données permet ainsi au web sémantique de se populariser et facilite la création de données liées en donnant accès à des jeux de données externes de manière gratuite.

2.1.2 Les quatre principes des Linked Data

Berners-Lee introduit le terme Linked Data dans un article publié en 2006. Il en définit quatre principes de base pour relier les données (Berners-Lee 2010) :

- Utiliser des URIs pour désigner les ressources
- Utiliser des HTTP URIs pour rendre les ressources déréférençables
- Fournir les données utiles pour chaque URI en utilisant des standards (RDF, SPARQL)
- Inclure des liens vers d'autres URIs pour favoriser la découverte

Avant 2006, de nombreux jeux de données n'étaient pas liés car ils ne respectaient pas l'un ou l'autre de ces principes. Le but de l'auteur est donc de communiquer une base technique sur laquelle chacun devrait s'appuyer pour créer des Linked Data interopérables, fonctionnelles et utilisables de manière optimale.

2.1.3 Les cinq étoiles des Linked Open Data

En 2010, Berners-Lee a complété ses quatre principes en y ajoutant le modèle des cinq étoiles des LOD (tableau 1).

Tableau 1: Les cinq étoiles des Linked Open Data

Les données doivent être...

★	disponibles sur le web (quel que soit le format), mais sous licence ouverte, pour de l'Open Data
★★	structurées pour le traitement automatique par des ordinateurs (un tableau Excel plutôt qu'un tableau scanné)
★★★	disponibles sous un format non-propriétaire (CSV plutôt qu'Excel)
★★★★	décrites selon les standards ouverts du W3C (RDF et SPARQL), pour que les gens puissent s'y référer
★★★★★	reliées à des données externes afin d'être contextualisées

(Berners-Lee 2010)

L'interopérabilité technique des Linked Data, décrite dans les quatre principes, est ainsi complétée par l'aspect d'ouverture des données. Ce schéma vise notamment à encourager les gouvernements à publier leurs données (Open Government Data).

2.1.4 Aspects techniques, normes, standards

Les définitions et schémas précédents sont basés sur quelques termes techniques spécifiques au web sémantique. Ce chapitre les explique brièvement.

2.1.4.1 URI et IRI

Tandis qu'un URL (Uniform Resource Locator) permet de localiser tout ce qui existe sur le web, un URI (Uniform Resource Identifier) est une séquence de caractères qui permet d'identifier, sur le web, tout ce qui existe (Gandon 2013, p. 11). Ceci inclut des ressources abstraites et des objets du monde réel.

Les URIs ont été standardisés au sein d'une spécification formelle en 1998. En 2005, une nouvelle spécification a été publiée : les IRIs, ou Internationalized Resource Identifiers. Un IRI peut être composé de caractères UNICODE, comme des signes japonais, alors que les URIs sont limités aux caractères ASCII (Dürst, Suignard 2005, chap. 1.1). Le W3C utilise le terme *IRI* dans ses recommandations officielles.

Les URIs et les IRIs décrivent les ressources indépendamment de leur format. Les URL sont plus spécifiques ; ils décrivent également la méthode d'accès au contenu

(Berners-Lee, Fielding, Masinter 2005, chap. 1.1.3). Ainsi, un document exprimé en format PDF et le même document exprimé en format Word pourront être dotés d'une IRI ou d'une URI similaire, mais auront en tous les cas deux URL différentes.

Créer du Linked Data requiert l'utilisation de HTTP URIs (c.f. chapitre 2.1.2) ou IRIs. Il s'agit d'identifiants utilisant le protocole HTTP afin de faire partie intégrante du web. Déréférencer un HTTP URI ou IRI signifie envoyer une requête HTTP afin d'obtenir une représentation de la ressource.

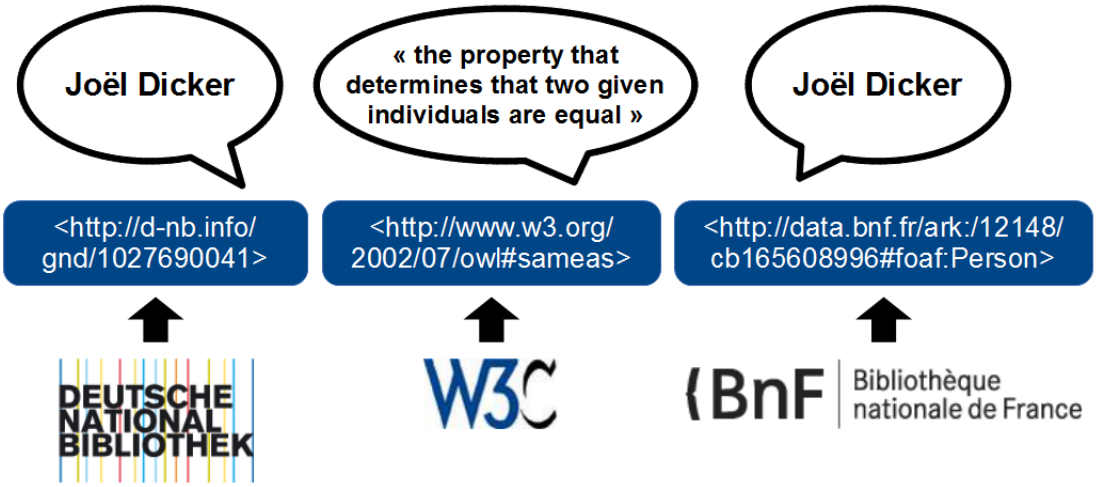
Alors que le *Web of documents* se base essentiellement sur des URL, le *Web of data* nécessite des URIs ou des IRIs. Cela lui permet d'identifier de façon universelle des ressources abstraites ou réelles autres que des documents présents sur le web, et d'établir des liens sémantiques entre elles

2.1.4.2 RDF[1]

> *« The Resource Description Framework (RDF) is a framework for expressing information about resources. Resources can be anything, including documents, people, physical objects, and abstract concepts. »* (W3C 2014a, chap. 1)

RDF est un modèle de données permettant de structurer l'information afin qu'elle puisse être traitée de manière automatique par des ordinateurs. Il peut également servir de modèle d'échange de données et permet de créer des liens entre diverses ressources, y compris des ressources provenant de jeux de données différents.

Figure 1: Triplet RDF



[1] Ce chapitre est basé sur le document *RDF 1.1 Primer* (W3C 2014a) publié par le W3C, contenant les spécifications officielles de RDF.

Une donnée RDF est exprimée sous la forme d'un triplet de type *sujet-prédicat-objet*. La figure 1 illustre cette structure : l'entité *Joël Dicker* de la Bibliothèque nationale allemande (DNB) est le sujet, la propriété *same as* est le prédicat, et l'entité *Joël Dicker* de la Bibliothèque nationale de France (BnF) est l'objet.

Le sujet et le prédicat d'un triplet RDF doivent être identifiés par des IRIs, alors que l'objet peut consister soit en un IRI, soit en du texte brut. Le modèle utilise des classes et des propriétés pour structurer les données. Le langage RDF Schema (RDFS) permet de définir ce type d'éléments. La précision d'un modèle peut être approfondie grâce à l'utilisation du Web Ontology Language (OWL).

RDF se veut indépendant de toute syntaxe. Des données RDF peuvent donc être exprimées en divers formats, appelés *sérialisations*. En voici les principales :

- Famille Turtle : N-Triples, Turtle, TriG and N-Quads
  N-Triples est la sérialisation la plus simple, exprimant RDF selon la structure *sujet-prédicat-objet* dans un document où chaque ligne correspond exactement à un triplet. Les trois autres sérialisations sont des extensions de la première, dont le but est soit d'améliorer la lisibilité du code, soit d'exprimer des informations impossibles à transcrire avec la syntaxe N-Triples.
- JSON-LD
  JSON (JavaScript Object Notation) est un format d'échange de données léger, facile à lire par l'œil humain et facilement traitable par ordinateur, s'exprimant selon des conventions auxquelles sont habitués les programmeurs (ECMA International 2013). JSON-LD (pour Linked Data) est une extension de JSON servant à exprimer du RDF.
- RDFa
  RDFa (RDF in attributes) a pour but d'intégrer des données RDF dans des documents HTML ou XML. Cela permet par exemple aux moteurs de recherche ou aux navigateurs web d'exploiter ces données structurées afin d'améliorer leurs services.
- RDF/XML
  Cette sérialisation permet d'exprimer des données RDF selon la syntaxe XML. Aux débuts du RDF, il s'agissait de la seule sérialisation existante. Elle est encore aujourd'hui la plus répandue et on la nomme parfois simplement RDF.

2.1.4.3 SPARQL

SPARQL (SPARQL Protocol and RDF Query Language) regroupe une série de spécifications définissant des langages et protocoles pour interroger et manipuler des données RDF, à l'instar de SQL pour les bases de données (W3C 2013b, chap. 1). Un *SPARQL endpoint* est un service web acceptant et répondant à des requêtes SPARQL (W3C 2013c).

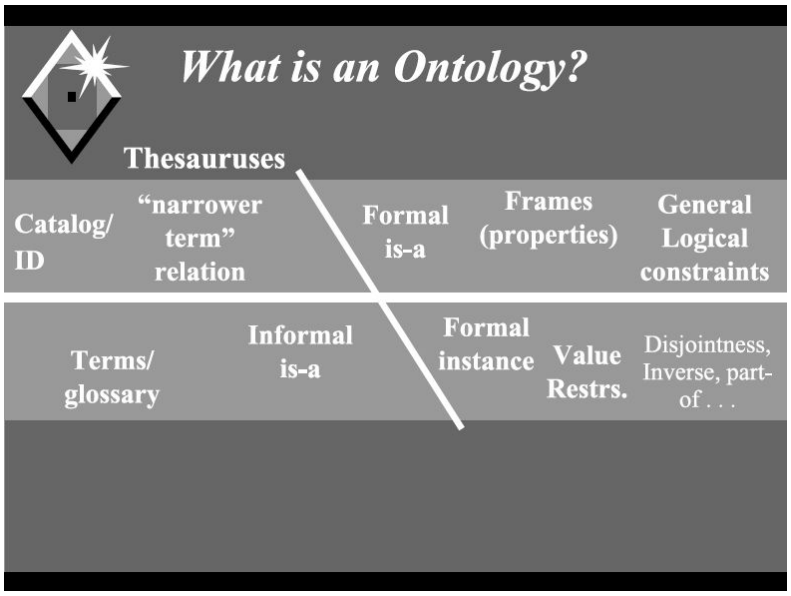
2.1.5 Ontologies et vocabulaires

Dans le contexte du web sémantique, les termes *ontologie* et *vocabulaire* sont souvent confondus. La littérature professionnelle n'est pas catégorique sur leurs définitions.

Browne et Jermey relèvent l'ambiguïté de ces concepts en fournissant deux définitions. Au sens strict, une ontologie fournit des relations précises définies au moyen de langages de représentation d'ontologies, comme RDFS ou OWL. Au sens large, cette définition inclut les taxonomies et les thésaurus (Browne, Jermey 2001, p. 94).

McGuinness établit un spectre des interprétations possibles du mot *ontologie* (figure 2), allant du plus simple vocabulaire contrôlé à un langage contenant des contraintes logiques telles que relations inverses, classes disjointes, etc. L'auteur distingue néanmoins les ontologies simples des ontologies structurées (séparées par une ligne de biais sur la figure 2).

Figure 2: Eventail d'interprétations du mot *ontologie*



(McGuinness 2003, p. 175)

De même, le W3C (2013d) ne fait pas de différence claire entre *ontologie* et *vocabulaire*, mais utilise par convention le premier pour des éléments de métadonnées plus complexes.

Dans le contexte spécifique des bibliothèques et du web sémantique, le *Library Linked Data Incubator Group* du W3C (2011, chap. 1) identifie trois types de ressources :

- Les éléments de métadonnées, tels que les termes Dublin Core, fournissent des classes et des propriétés (ou relations) permettant de décrire des

ressources. Ils sont définis au moyen de langages comme RDFS ou OWL. Les mots *ontologie* et *vocabulaire RDF* sont couramment utilisés pour désigner des éléments de métadonnées.

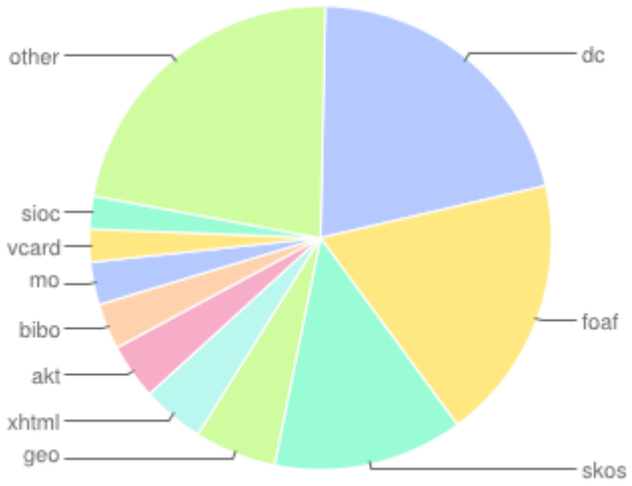
- Les référentiels ou vocabulaires de valeurs sont constitués de valeurs contenues au sein d'éléments de métadonnées. Généralement, les ressources décrites concernent des domaines particuliers, tel que des sujets, des personnes ou des lieux, et possèdent des IRIs. En bibliothèque, ces référentiels correspondent aux vocabulaires contrôlés et thésaurus, tels que les répertoires d'autorités RAMEAU[2] et LCSH[3].
- Les jeux de données sont des ensembles de données structurées, correspondant en bibliothèque aux notices bibliographiques par exemple. Un tel jeu de données sera composé d'éléments de métadonnées dont les valeurs sont des littéraux ou proviennent de divers référentiels de données.

La distinction ci-dessus sera utilisée dans ce travail. Les termes *ontologie* et *vocabulaire RDF* seront donc associés indifféremment aux éléments de métadonnées.

2.1.5.1 Ontologies les plus utilisées sur le web

La figure 3 montre l'utilisation des ontologies sur le web. Les trois plus utilisées sont clairement DC, FOAF et SKOS, décrites plus en détail dans le tableau 4.

Figure 3: Ontologies les plus utilisées sur le web



(Jentzsch, Cyganiak, Bizer 2011)

Les ontologies ci-dessus concernent des domaines particuliers : localisation géographique, personnes, musique, etc. Beaucoup servent à décrire des ressources bibliographiques. Néanmoins, d'autres ontologies spécialisées ont été développées pour pouvoir décrire des documents plus finement.

[2] *Répertoire d'autorité-matière encyclopédique et alphabétique unifié*, développé par la BnF
[3] *Library of Congress Subject Headings*, l'équivalent américain de RAMEAU

2.2 Les métadonnées des bibliothèques

Les métadonnées des bibliothèques décrivent primairement les ressources documentaires qu'elles gèrent. De façon traditionnelle, ces métadonnées sont enregistrées sous forme de notices bibliographiques dans des catalogues. Par ailleurs, les bibliothèques collectent et créent également des données sur d'autres types d'éléments liés aux documents : personnes, organisations, événements, lieux, concepts, etc. Ces données sont enregistrées au sein de notices dites d'autorité.

Le rôle des notices d'autorité est de gérer les points d'accès aux notices bibliographiques. Il s'agit des éléments qui apparaissent dans un index alphabétique accessible et visible par l'usager, tel que l'index des sujets ou des auteurs. Une notice d'autorité se compose d'une vedette – la forme retenue d'un nom pour désigner une entité et le point d'accès autorisé – ainsi que de formes rejetées ou alternatives. Un usager effectuant une recherche à partir de l'une des formes rejetées trouvera ainsi, à travers la vedette, l'ensemble des notices bibliographiques associées (Willer 2009, p. 16).

Les fonctionnalités telles que celle-ci (recherche alphabétique, par auteur, etc.) sont proposées dans les catalogues de bibliothèques traditionnels. Or leur utilité est remise en question par l'avènement du web. Les utilisateurs effectuent aujourd'hui des requêtes libres dans les grands moteurs de recherche et visent un accès instantané aux documents en ligne. Les catalogues des bibliothèques, accessibles uniquement au moyen d'un formulaire, ne sont pas indexés par les moteurs de recherche et restent dans le web profond. En effet, dans la plupart des bibliothèques, les données des catalogues sont gérées, échangées et enregistrées dans un format développé bien avant le web et peu compatible avec celui-ci : le format MARC (c.f. chapitre 2.2.1).

Une réforme est donc en cours. Bermès (2013, chap. 1.3) parle de « réconcilier les normes des bibliothèques avec celles du web », afin de décloisonner les données et les rendre plus interopérables. Cet objectif peut être atteint grâce au web des données.

2.2.1 Le format MARC et ses dérivés

> *« The MARC formats are standards for the representation and communication of bibliographic and related information in machine-readable form. »*
>
> (Library of Congress 2014a)

MARC est l'abréviation de *MAchine-Readable Cataloging*. Le format MARC fournit la base technique par laquelle les ordinateurs échangent, utilisent et interprètent les données bibliographiques. Il a été développé vers la fin des années 1960 à la

Bibliothèque du Congrès (LOC) aux Etats-Unis. Au fil du temps, MARC fut repris par les bibliothèques nationales et adapté aux besoins locaux, se déclinant en de nombreux formats dérivés tels que CANMARK, UKMARC ou USMARC. Face aux problèmes d'interopérabilité, un format d'échange a été créé par la Fédération internationale des associations et institutions de bibliothèques (IFLA) sous le nom d'UNIMARC. En Amérique du Nord naissait en parallèle le standard MARC21, résultat de la fusion de plusieurs formats dérivés. Tandis que ce dernier acquérait une portée internationale grâce à son adoption par la LOC, UNIMARC s'est imposé seulement dans certaines bibliothèques, notamment en France (Fürste 2011, p. 18, 21; McCallum 2010).

Une notice MARC se définit selon trois éléments (Bermès 2013) :

- La structure du format :
  La norme ISO 2709 définit cette structure en quatre parties : guide ou label de notice, répertoire, zones de données, séparateur de notice. Les zones de données contiennent les données elles-mêmes, un séparateur de zones et, de manière facultative, un indicateur et un identificateur (AFNOR 1987, p. 574).
- Les éléments de données :
  Il s'agit des codes de zones, composés de trois chiffres, ainsi que des codes de sous-zones, commençant par un caractère distinctif comme par exemple "$". Ces codes sont définis par le type de format MARC.
- Les données elles-mêmes :
  Le contenu des données ne dépend pas directement du format, mais des règles établies par les organisations (règles AACR2 ou RDA dans le cas des bibliothèques).

Un exemple commenté de cette structure se trouve dans l'annexe 1. MARC21 et UNIMARC sont les deux formats reconnus officiellement par l'IFLA (Leresche 2004).

2.2.1.1 Les problèmes du format MARC

Après l'avoir utilisé pendant près de 50 ans, les bibliothèques reprochent aujourd'hui au format MARC un certain nombre de limites (Fürste 2011, p. 26-30; Tennant 2002) :

- Gestion très limitée des relations entre notices. Les liens sont créés plus souvent sur la base de chaînes de caractères textuelles que d'identifiants uniques, ce qui exige un entretien plus important des données.
- Identification des éléments peu harmonisée. Trop peu de champs sont contrôlés, ce qui limite le traitement automatique des données. Par exemple, si les sujets étaient décrits par des identifiants plutôt que par des mots, il serait plus aisé de construire un système multilingue.
- Marginalisation technique. Seules les bibliothèques utilisent MARC tandis que les autres acteurs de l'information sont passés à des standards basés sur XML ou d'autres formats.
- Format difficilement extensible et adaptable aux nouvelles données. De par

son ancienneté exceptionnelle, « son adaptation à de nouveaux besoins a conduit à le complexifier et le risque d'introduire une divergence avec les données existantes freine les évolutions majeures » (Bermès 2013, chap. 2.1).

Ces divers problèmes mettent en évidence le fait que MARC a été développé sur le modèle des catalogues sur fiches : nombre d'index limité, champs trop textuels, liens entre notices rares et difficiles, etc. Malgré cela, ce format est toujours utilisé.

> *« There's understandably a great reluctance to tackle a change that will have such a wide-ranging effect on our everyday library operations. » (Coyle 2006)*

Abandonner MARC signifierait pour la communauté des bibliothèques de reconstruire toute l'infrastructure de ses systèmes d'information, car celle-ci repose aujourd'hui entièrement sur MARC. Face à cette problématique, les bibliothèques cherchent des évolutions pour la gestion de leurs données.

2.2.2 Vers de nouveaux standards

Les années 2000 voient les bibliothèques en pleine transition sur plusieurs plans quant à la gestion de leurs données bibliographiques. D'une part, elles se créent avec FRBR (c.f. chapitre 2.2.2.1) un nouveau modèle conceptuel pour l'organisation des données, et tentent de l'appliquer à leurs infrastructures actuelles. D'autre part, elles adaptent leurs règles de catalogage à l'environnement numérique en pleine expansion, en introduisant RDA (c.f. chapitre 2.2.2.2). Enfin, les bibliothèques se cherchent, au moyen de l'initiative BIBFRAME (c.f. chapitre 2.2.2.3), une alternative au format MARC, qui ne répond plus à leurs besoins.

Or, ces trois facettes sont étroitement étudiées lors de toute réflexion liée au web sémantique, car elles représentent des facteurs essentiels du décloisonnement des données de bibliothèques.

2.2.2.1 FRBR[4]

Les *Fonctionnalités requises des notices bibliographiques*, abrégées FRBR (Functional requirements for bibliographic records), se concrétisent sous la forme d'un modèle conceptuel pour les données des bibliothèques. Ce modèle se base sur les fonctionnalités des notices et les besoins des utilisateurs. FRBR a été publié pour la première fois en 1997 par l'IFLA, puis actualisé et complété par deux nouvelles recommandations : Fonctionnalités requises des données d'autorité (FRAD) et des données d'autorité matière (FRSAD). Ensemble, elles forment un seul modèle global

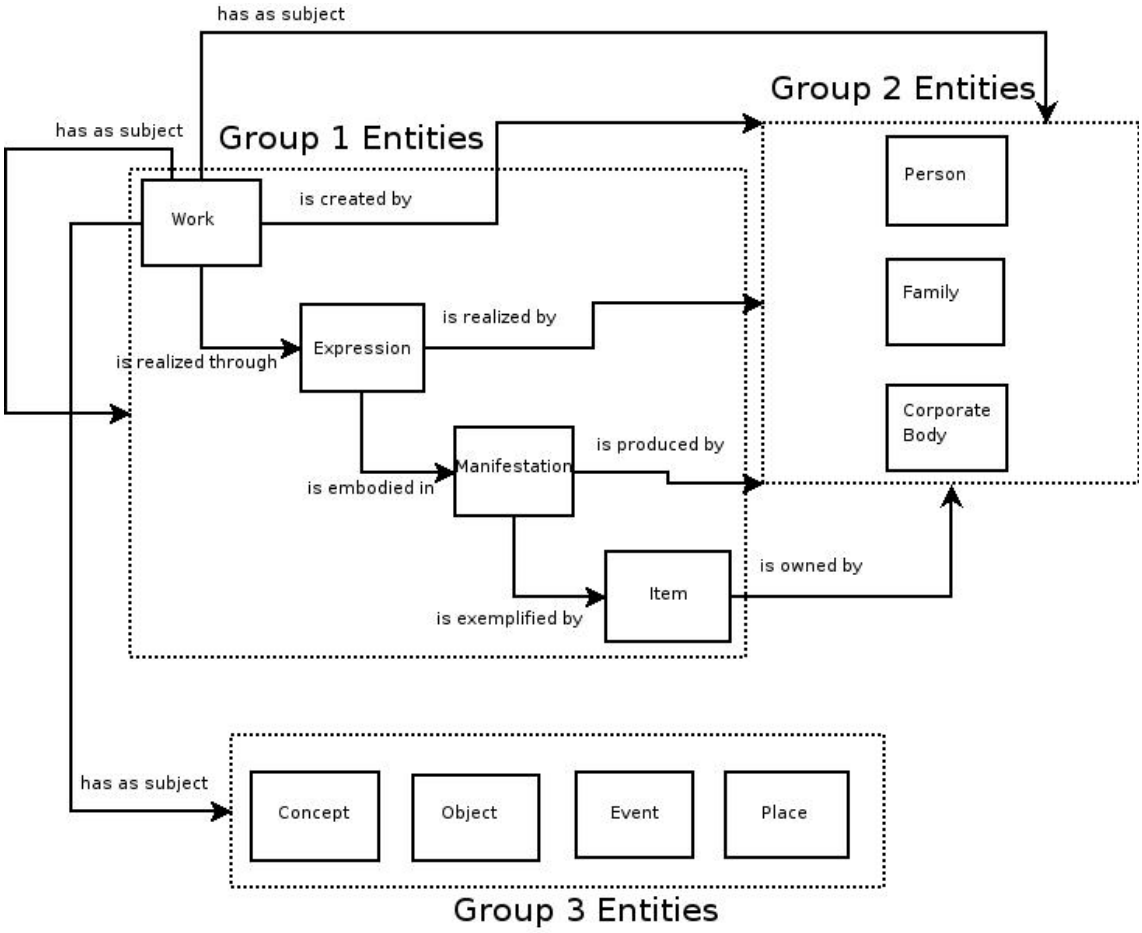
[4] Ce chapitre est basé sur les rapports finaux de l'IFLA concernant FRBR (IFLA 2012a), FRAD (IFLA 2010) et FRSAD (IFLA 2012b).

de type entité-relation, comprenant trois groupes d'entités :

- Groupe 1 : Œuvre, Expression, Manifestation, Item (entités décrites dans FRBR)
- Groupe 2 : Personne, Famille, Collectivité (décrites dans FRAD)
- Groupe 3 : Concept, Objet, Evénement, Lieu (décrites dans FRSAD)

Ces entités peuvent être reliées entre elles de plusieurs manières (figure 4) afin de former des notices bibliographiques cohérentes.

Figure 4: Relations entre les entités FRBR



(Denton 2006)

On parle souvent en bibliothèque de *FRBRiser* les données. Cela signifie essentiellement adapter les notices bibliographiques actuelles au schéma œuvres-expression-manifestation-item. Concrètement, une notice MARC telle qu'elle se présente dans les catalogues traditionnels correspond à une manifestation. Le travail consiste alors à créer, sur cette base, des notices *mères* d'expression et d'œuvre.

Ainsi, une œuvre regroupera en principe plusieurs expressions, qui chacune regroupera plusieurs manifestations. Les informations d'items sont généralement séparées de la notice MARC dans l'utilisation actuelle que l'on fait de ce format. Cette hiérarchisation de la description bibliographique est en fait la principale innovation de FRBR, notamment avec l'entité œuvre.

Ce modèle sied particulièrement au web sémantique, car il identifie formellement des entités qui ne sont pas uniquement d'ordre bibliographique (personnes, événements, lieux, etc.). Or, ces entités sont également présentes dans d'autres jeux de données et peuvent être reliées. L'œuvre est l'exemple type de cet argument. Une manifestation représente un objet réel bien particulier qui n'existe que dans certaines bases de données. A l'inverse, une œuvre, par exemple le roman *Les dix petits nègres* d'Agatha Christie, peut se décliner en films, pièces de théâtre, jeux vidéos, etc. Chaque déclinaison se trouve probablement décrite en détail dans une base de données spécialisée et peut être reliée aux autres grâce à la notion d'œuvre. Pour l'intégration du modèle au web sémantique, une ontologie FRBR a été développée par l'IFLA.

Toutefois, malgré sa première publication il y a dix-sept ans déjà, FRBR peine à s'imposer. En tant que modèle entité-relation, son implémentation est rendue difficile par le format MARC, dont l'adaptation est limitée (Le Pape 2014a) et qui gère difficilement les liens entre les notices. Les bibliothèques implémentent alors FRBR, mais souvent en aval de la saisie des données, lors de l'affichage public sur le catalogue en ligne, au moyen de logiciels regroupant automatiquement des notices de manifestations au sein d'expressions et d'œuvres. Cela signifie que ces regroupements sont effectués à la volée, lorsque l'utilisateur a soumis une requête. En outre, passer au modèle FRBR implique également de transformer de manière automatique les très nombreuses notices classiques déjà existantes, ce qui engendre forcément des erreurs. Néanmoins, ce modèle atteint en ce moment une nouvelle visibilité grâce au déploiement de RDA.

2.2.2.2 RDA[5]

RDA (abréviation de Resource Description and Access) est un standard de description de ressources destiné à remplacer les règles de catalogage AACR2[6]. Publié en 2010, il a été développé par le Joint Steering Committee for Development of RDA, regroupant des bibliothèques anglo-saxonnes.

[5] Ce chapitre est basé sur les deux sources suivantes : (Kiorgaard 2009; BnF 2013a)
[6] Anglo-American Cataloguing Rules, 2e édition

RDA concerne aussi bien les notices bibliographiques que les notices d'autorité. Il a été conçu pour satisfaire les besoins issus de l'environnement numérique croissant. L'une de ses innovations est l'identification précise des types de document, qui était devenue confuse dans les AACR2 à cause de l'arrivée du web et de nouveaux supports. RDA distingue ainsi trois facettes des documents :

- Le type de contenu : texte, image fixe, image cartographique, musique imprimée, etc.
- Le type de média (correspondant aux moyens nécessaires pour accéder au contenu de la ressource) : sans média, audio, vidéo, ordinateur, etc.
- Le type de support : support audio (cassette, disque, etc.), support informatique (disque, en ligne, etc.), support sans média (volume, rouleau, carte, etc.), etc.

Une autre nouveauté importante de RDA consiste en son intégration du modèle FRBR, de sa terminologie et de ses relations entre entités. Les nouvelles règles mettent l'accent sur les différents types de liens qui peuvent exister entre notices et sont optimisées pour le web sémantique. Une ontologie RDA ainsi que plusieurs référentiels de valeurs ont été publiés en ligne.

Par ailleurs, RDA fait preuve d'une certaine flexibilité dans ses règles et contraintes, car il ambitionne d'être utilisé dans d'autres contextes que celui des bibliothèques, comme les archives ou les musées. Ainsi, des métadonnées d'origines différentes s'intégreront plus facilement.

Né aux Etats-Unis, RDA y est appliqué depuis 2013. En Europe, les pays germanophones (Allemagne, Autriche, Suisse alémanique) ont débuté en commun le travail d'implémentation des règles, qui se terminera en principe à la fin 2015 (Behrens, Schaffner 2014). La France n'a pas considéré RDA comme assez satisfaisant et abouti pour l'adopter, mais pense le faire dans le futur (BnF 2013b).

2.2.2.3 BIBFRAME[7]

L'initiative BIBFRAME (abréviation de Bibliographic Framework), lancée en mai 2011 et menée par la LOC, veut créer un nouveau standard de représentation et d'échange de données bibliographiques, adapté au web. Le but est d'intégrer les données des bibliothèques au web sémantique et de remplacer le format MARC.

Le résultat de cette initiative est un méta-modèle défini en RDF et conçu pour être compatible avec plusieurs standards de description existants (RDA, VRA[8], etc.). Il est

[7] Ce chapitre est basé sur les deux sources suivantes : (Library of Congress 2012, 2014b)

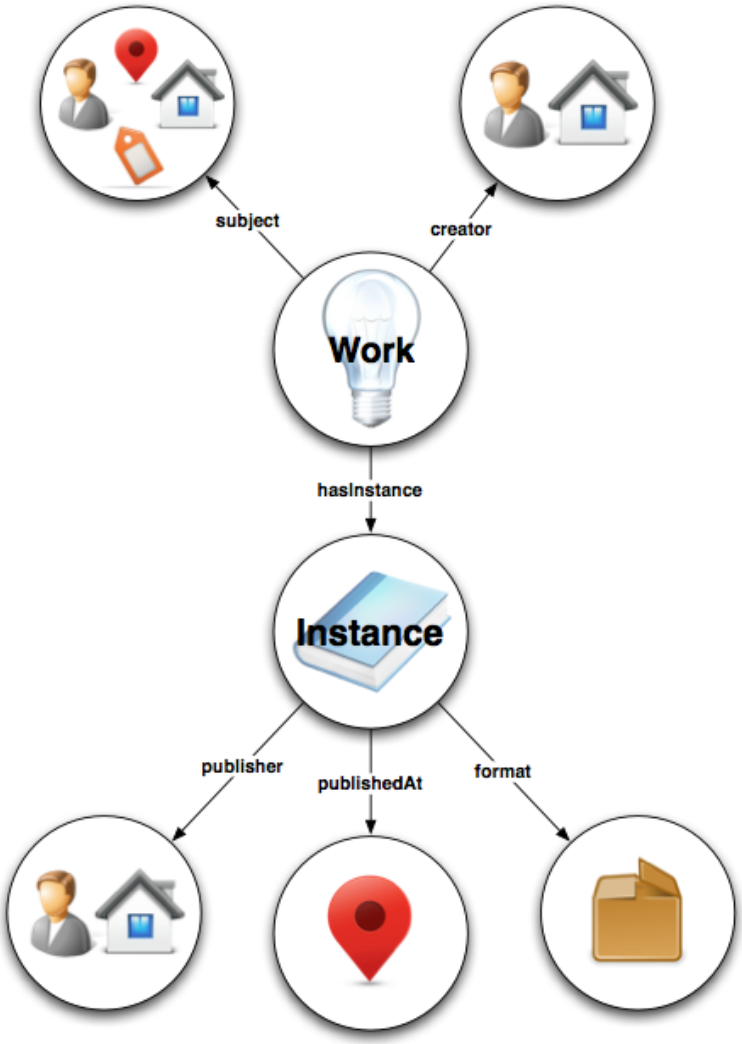
[8] Abr. de Visual Resources Association, un standard pour la description d'images et d'œuvres d'art

intentionnellement spécifié de manière simplifiée (contraintes logiques plus souples) pour être flexible et répondre aux besoins des diverses communautés culturelles (bibliothèques, archives, musées, etc.).

> *« [...] BIBFRAME is intended to be both "rule agnostic » (i.e., not tied to a particular cataloging code) and "model agnostic" (i.e., flexible enough to accommodate both "flat" record-based as well as highly interlinked FRBRized data). »* (Svensson 2013, p. 9)

L'idée consiste à créer des *BIBFRAME community profiles* qui établiront une compatibilité – au moyen de correspondances – avec d'autres modèles, tels que FRBR (annexe 2).

Figure 5: Modèle BIBFRAME



Le modèle BIBFRAME se compose essentiellement des quatre classes suivantes : Œuvre, Instance, Autorité, Annotation (figure 5). La classe Annotation n'apparaît pas

dans la figure 5 ; elle comprend toutes sortes de ressources liées aux trois autres classes, comme par exemple des informations d'item, des vignettes de couverture, des critiques, etc. Le modèle et son vocabulaire RDF sont encore en pleine phase de développement.

2.3 Réalisations

C'est en 2008 que les premières données de bibliothèques sont publiées en LOD. La LOC met à disposition le LCSH tandis que le réseau Libris, géré par la Bibliothèque royale de Suède, fournit l'ensemble de son catalogue en RDF (Pohl, Danowski 2013, p. 13).

Peu à peu la tendance se répand et d'autres institutions s'y mettent, notamment des bibliothèques nationales (Allemagne, Espagne, France, Grande-Bretagne, etc.) et universitaires (HBZ[9] en Allemagne, SUDOC[10] en France, etc.). En 2014, OCLC[11] a également publié sa base de données bibliographiques *WorldCat* en LOD (OCLC 2014a).

Selon l'optique de Libris, la simple publication ne suffit plus. Il faut à présent utiliser les données, les rendre utiles, grâce à des interfaces. Au modèle des cinq étoiles des LOD introduit par Tim Berners-Lee, Martin Malmsten (2013, p. 34-35) se permet ainsi d'en ajouter une sixième : l'expérience utilisateur.

Dans ce sens, quelques bibliothèques ont développé des applications exploitant la plus-value des LOD. En France par exemple, la BnF (2014b) a créé le portail *data.bnf.fr* donnant accès à diverses ressources, issues du catalogue général, de la base de données des manuscrits et de la bibliothèque digitale Gallica, reliées entre elles et à des référentiels externes. A Paris, le Centre Pompidou (2014), regroupant entre autres une bibliothèque et un musée d'art, a totalement repensé son site web en fonction des Linked Data. Il fournit à présent un accès centralisé à des ressources variées – allant de la notice bibliographique au dossier pédagogique – et reliées entre elles. Enfin, la plate-forme web finlandaise Kulttuurisampo (2014) met à disposition divers types de ressources (livres, objets de musée, vidéos, etc.) en LOD, assorties de fonctionnalités de découverte innovantes.

[9] Abr. de Hochschulbibliothekszentrum (Centre des bibliothèques des hautes écoles de Rhénanie-du-Nord-Westphalie)

[10] Abr. de Système universitaire de documentation (catalogue des institutions de l'enseignement supérieur et de la recherche françaises)

[11] Abr. de Online Computer Library Center, une coopérative internationale à but non lucratif dans le domaine des bibliothèques.

3. Contexte institutionnel[12]

Le Réseau des bibliothèques de Suisse occidentale, ou RERO (acronyme de réseau romand), regroupe principalement des bibliothèques universitaires, patrimoniales, publiques, scolaires et spécialisées. Ces bibliothèques proviennent de communes et de cantons majoritairement francophones, ainsi que d'institutions de la Confédération dans le domaine du droit (Institut suisse de droit comparé, tribunaux fédéraux).

Fondé en 1985 sous les auspices de la Conférence universitaire de Suisse occidentale, le réseau est depuis 2009 placé sous la haute surveillance de la Conférence intercantonale de l'instruction publique de la Suisse romande et du Tessin. RERO est devenu aujourd'hui l'un des principaux réseaux de bibliothèques en Suisse. Il est subdivisé en six sous-réseaux (Fribourg, Genève, Jura/Neuchâtel, Valais, Vaud, institutions fédérales), chacun représenté par un coordinateur local. La coordination et la gestion des activités du réseau se font à la Centrale RERO, basée à Martigny.

En tant qu'instrument de politique documentaire soutenu par les autorités politiques, académiques et culturelles de Suisse occidentale, RERO a notamment pour mission de proposer à ses bibliothèques membres des outils et des services, de mettre en valeur leurs ressources grâce notamment à un catalogue collectif, de coordonner les pratiques professionnelles, d'établir des coopérations avec ses membres et avec des organes extérieurs, d'optimiser la gestion des coûts, et de participer à la pérennisation du patrimoine intellectuel.

3.1 Premiers pas réalisés vers les Linked Open Data

La gestion des données bibliographiques et le web sémantique représentent l'un des axes du plan stratégique 2013-2017 de RERO. Cet axe se décline en quatre sous-objectifs (RERO 2012, p. 2) :

- Introduire les nouveaux standards mondiaux (par ex. RDA).
- Accroître la plus-value des métadonnées de RERO par la sémantisation.
- Développer des services web (API) pour la fourniture de métadonnées sémantisées aux sites.
- Optimiser la visibilité des données de RERO.

[12] Chapitre inspiré du rapport d'activités 2013 ainsi que du site web RERO (RERO 2014a, 2014b).

Cette stratégie inclut la publication des données en LOD. Dans cette optique, RERO a déjà effectué plusieurs actions préparatoires (Moreira 2013, p. 27):

- 2010 : adhésion au Virtual International Authority File (VIAF).
- 2011 : adoption du vocabulaire d'indexation RAMEAU de la BnF, qui a déjà été publié en LOD et relié à d'autres vocabulaires.
- 2012 : mise en place d'une politique d'encouragement à la création de notices d'autorité auteur-collectivité, qui sont la clé de voûte des LOD en bibliothèque.
- 2012 : attribution d'URIs aux données, pré-requis pour la conversion en LOD.
- 2014 : ouverture des métadonnées, avec leur mise dans le domaine public selon les termes CC0 (Creative Commons Zero) (RERO 2014c).

La prochaine étape consiste donc à modéliser et à transformer les données selon le modèle RDF. C'est l'objet de ce travail.

4. Méthodologie générale

Le projet de transformation des données de RERO en LOD s'est déroulé en collaboration avec un spécialiste de la centrale RERO, qui a participé aux discussions et aux choix effectués, en relayant la discussion de certains aspects à d'autres spécialistes de son entourage. Cette personne intervenait ensuite concrètement sur les données avec des outils informatiques, dont une partie était développée en interne pour ce projet. Les tâches des deux participants étaient étroitement liées. La thématique du web sémantique, relativement récente et peu connue, a donné à cette entreprise un côté expérimental.

Le projet RERO ne s'est pas déroulé en une succession d'étapes linéaires, mais plutôt en suivant un cycle de phases qui se sont répétées et complétées. Ce processus, illustré dans la figure 6, s'inspire des meilleures pratiques pour la publication de données gouvernementales en LOD (Villazón-Terrazas et al. 2011; W3C 2014b, chap. 1). Il a été adapté au contexte des bibliothèques selon les besoins particuliers de RERO et selon les recommandations émises par des institutions ayant conduit des

Figure 6: Processus de développement

projets similaires, notamment la Bibliothèque nationale d'Espagne (BNE) (Vila-Suero, Gómez-Pérez 2013, p. 580) et la Bibliothèque royale de Suède (Malmsten 2009).

La méthode adoptée a permis un travail concerté entre les deux participants, structuré par des rencontres toutes les deux semaines environ. Ceci avait pour but d'encourager l'obtention rapide de résultats concrets en raison de contraintes de temps, la durée du projet étant de six mois seulement. L'un des objectifs fixés était, dans cette optique, de parvenir à la création de données RDF au plus tôt dans le cycle du projet. Le perfectionnement des données créées se ferait progressivement, et la publication n'interviendrait qu'en fin de projet.

Cette manière de faire correspond tout à fait au processus *en cycle* de la figure 6, qui permet d'effectuer lors d'une première itération des actions sommaires à des fins de test, et de les compléter peu à peu lors des itérations suivantes. Cette approche suit également la devise « a “data first” approach is better than “perfect metadata first” » préconisée par Martin Malmsten (2009, chap. 1), responsable du développement et design de logiciels à la Bibliothèque royale de Suède. Une fois le projet terminé, le cycle exprime alors la nécessité d'une maintenance continuelle des résultats produits (applications, dumps[13], etc.) suivant ce processus.

Pour les premiers essais de conversion, un échantillon représentatif de l'ensemble des données du catalogue collectif a été sélectionné, afin de pouvoir tester et valider les actions effectuées.

Vu l'étendue du travail, une priorisation des tâches s'est révélée indispensable et certains aspects moins importants ont ainsi dû être mis de côté.

Ce mémoire et la méthodologie présentée se concentrent sur l'aspect qualitatif et bibliothéconomique de la transformation des données plutôt que sur l'aspect technique. Les étapes du processus illustré dans la figure 6 et les résultats intermédiaires sont décrits en détail dans le chapitre suivant.

[13] Les dumps sont des fichiers de données créés pour être téléchargés et réutilisés.

5. Processus et résultats

5.1 Revue de la littérature

Une importante partie du travail consiste à s'informer des meilleures pratiques d'autres bibliothèques ou acteurs concernant le web sémantique. Dans ce projet, cette analyse a pris la forme de nombreuses comparaisons des données de bibliothèques déjà publiées en RDF[14]. En effet, la normalisation est encore faible à ce jour et l'adoption des pratiques des autres contribue à la création de standards de facto.

La revue de la littérature inclut également une veille attentive sur les nouveautés dans ce domaine extrêmement évolutif. A titre d'exemple, plusieurs annonces majeures ont été faites au début de l'année 2014 :

- Janvier : publication de la nouvelle ontologie RDA
- Avril : publication des données OCLC en Linked Data
- Mai : mise à disposition d'un éditeur BIBFRAME par la LOC

5.2 Analyse des données

L'analyse des données internes (format, quantité, qualité, types d'entités, etc.) est un prérequis pour leur transformation en LOD (Vila-Suero, Gómez-Pérez 2013, p. 581).

Les données de RERO à transformer en LOD sont celles du catalogue collectif et de la bibliothèque numérique RERO DOC.

5.2.1 Données du catalogue collectif

Au 7 août 2014, le catalogue collectif comprenait 6'496'025 notices bibliographiques et 14'178'611 notices d'exemplaires, formant ainsi l'un des deux plus importants catalogues de Suisse. Ces données sont complétées par des notices d'autorité :

- 369'325 notices d'autorité sujet
- 217'910 notices d'autorité auteur
- 194'736 notices d'autorité classification

Dans ce projet, seules les notices bibliographiques et les notices d'autorité auteur et sujet ont été converties en LOD[15].

[14] Les données des bibliothèques/catalogues suivants ont été comparées : Bibliographie nationale britannique, LOC, Bibliothèque nationale de la Diète (Japon), BNE, BnF, DNB, HBZ, Libris, SUDOC.

[15] Les notices d'autorité classification n'ont pas été prises en compte, car elles ne forment pas une unique classification d'envergure, mais plusieurs petites, créées selon les besoins locaux des bibliothèques.

Toutes les données du catalogue collectif sont enregistrées en format MARC21 et gérées par le logiciel de gestion de bibliothèque Virtua, de la société VTLS. Elles sont saisies selon les règles de catalogage AACR2.

Afin de choisir les éléments de données à transformer, et également dans le but de prioriser les éléments les plus importants, des statistiques sur la structure des notices étaient nécessaires. Ainsi, la fréquence d'apparition des zones MARC21 dans les notices a été relevée sur la base d'un échantillon (tableau 2).

Tableau 2: Les 20 zones MARC21 les plus fréquentes dans RERO

Fréquence d'apparition	Zone MARC et son intitulé
100 %	001 numéro de contrôle
100 %	003 identité du numéro de contrôle
100 %	005 date et heure de la dernière mise à jour
100 %	008 éléments de données de longueur fixe
100 %	035 numéro de contrôle du système
100 %	039 numéro d'opérateur
100 %	040 source du catalogage
100 %	245 zone du titre et de la mention de responsabilité
99 %	072 code de sujet
90 %	260 adresse bibliographique
88 %	300 collation
87 %	999 mémoire des liens hiérarchiques dans l'arborescence
71 %	100 entrée principale (nom de personne)
45 %	700 entrée secondaire (nom de personne)
44 %	957 localisation (bibliothèques genevoises)
37 %	972 localisation (bibliothèques vaudoises)
36 %	500 note
34 %	956 localisation (bibliothèques fribourgeoises)
32 %	650 matières (nom commun)
29 %	490 mention de collection

Les zones les plus courantes ne contiennent pas forcément les données que l'on souhaite publier ; il s'agit principalement de zones de gestion dont la présence est obligatoire au sein de chaque notice. Néanmoins, cela donne un premier aperçu quantitatif du jeu de données. Pour compléter cette statistique, deux aspects supplémentaires ont été analysés.

Le premier est le niveau de contrôle du contenu des zones. Si le contrôle est fort, il est plus facile de publier le contenu sous forme de liens vers d'autres notices RERO ou vers des données externes, issues de référentiels. S'il n'y a qu'un faible contrôle, comme dans les zones où l'on peut insérer librement du texte, la génération de liens est plus difficile. Les types de contrôle suivants ont été identifiés dans les différentes zones :

- Identifiant international (ISBN, ISSN, etc.)
- Identifiant interne (numéro de contrôle, etc.)
- URL (accès électronique)
- Données générées automatiquement par Virtua (date de mise à jour, etc.)
- Vocabulaire contrôlé par Virtua (type de document, pays de publication, etc.)
  Il s'agit de valeurs que le catalogueur doit choisir dans un menu déroulant du logiciel. Il ne peut y avoir d'erreur de saisie.
- Vocabulaire contrôlé, mais pas par Virtua (code de langue, code de localisation, etc.)
  Il s'agit de valeurs provenant de référentiels locaux ou internationaux, mais dont la saisie n'est pas contrôlée par Virtua. Il peut donc y avoir des erreurs. Pour certaines zones, des mécanismes automatiques développés par RERO vérifient quotidiennement les valeurs saisies dans la journée et envoient des notifications aux catalogueurs en cas de détection d'erreurs.
- Vedettes (auteur, sujet, liaison avec une notice supérieure, etc.)
  Les vedettes contiennent des liens vers d'autres notices RERO, basés sur la chaîne de caractères contenue dans la zone. Une simple faute de frappe de la part du catalogueur suffit à créer un lien corrompu.

Le second critère de priorisation est la pertinence du contenu de la zone pour la publication et l'identification du document. Cet aspect est qualitatif et a été évalué sans méthode formelle, selon une échelle de un à trois.

Cette analyse a fourni une vue d'ensemble des données et a permis d'élaborer leur transformation de manière progressive, en se concentrant d'abord sur le plus important et ensuite, si le temps le permettait, sur les éléments secondaires.

Seules les notices bibliographiques du catalogue collectif ont été étudiées de manière aussi approfondie, car ce sont de loin les plus complexes. Les notices d'autorité auteur et sujet, plus simples, n'ont pas nécessité une analyse aussi poussée ; elles ont été converties dans leur intégralité.

5.2.2 Données de RERO DOC

En juillet 2014, RERO DOC contient plus de 27'000 documents. Les données sont enregistrées au format MARC21, mais selon un ensemble de règles simplifié, comparé

à celui qui régit le catalogue collectif RERO.

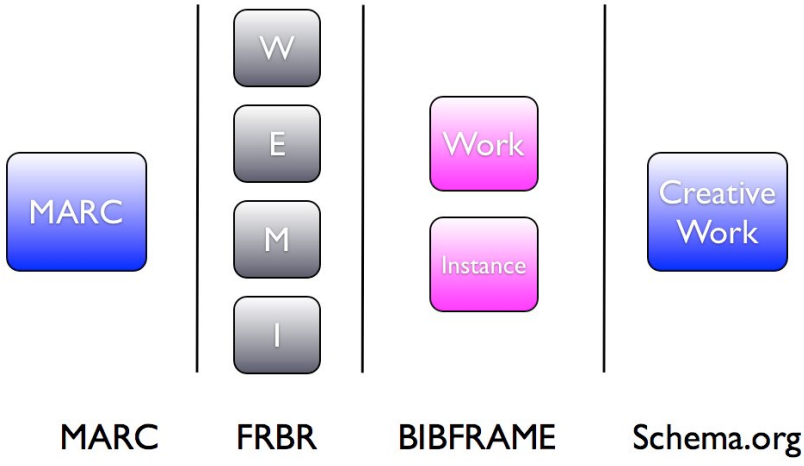
Leur analyse a été plus facile. Les données du catalogue collectif avaient été étudiées en premier et des règles de conversion existaient déjà pour la plupart des zones MARC, qui étaient équivalentes. Une analyse approfondie n'était donc plus nécessaire pour RERO DOC : la majeure partie du travail effectué sur le catalogue collectif a pu être reprise et adaptée.

5.3 Modélisation

5.3.1 Choix d'un modèle

Avant d'élaborer son propre modèle de données, il convient d'étudier les standards existants pour le domaine des bibliothèques, ainsi que les pratiques des autres institutions. Le chapitre 2.2.2 a introduit FRBR développé par l'IFLA, ainsi que BIBFRAME de la LOC. Ces deux modèles se distinguent de la pratique actuelle par leur hiérarchisation des notices bibliographiques (figure 7).

Figure 7: Comparaison de quatre modèles de données



(Coyle 2013)

Le modèle FRBR sépare une description bibliographique en quatre notices. BIBFRAME prévoit deux niveaux seulement, mais inclut les informations d'item au sein d'un troisième niveau grâce aux annotations. Le modèle basé sur le format MARC tel qu'il a été conçu à l'origine est dit *plat* (*flat* en anglais) ; il inclut toutes les données en une seule notice (Library of Congress 2012, p. 7), mais l'utilisation que les bibliothèques en font aujourd'hui sépare les informations d'item dans un niveau supplémentaire. Schema.org, en tant que modèle généraliste, n'a pas été considéré en

détail dans ce travail. Sa structure est néanmoins similaire à celle du modèle basé sur MARC.

Ces divers modèles, plus ou moins standardisés et adoptés par la communauté des bibliothèques, présentent chacun des avantages et des inconvénients. Par exemple, le morcellement d'une description en plusieurs notices représente d'un côté une distinction intellectuelle importante pour l'utilisateur des données, mais posera d'un autre côté des difficultés d'interaction avec les ressources web externes décrites en une seule entité (Coyle 2013). Ces difficultés se ressentent dans la mise en œuvre même du modèle FRBR par exemple, où la distinction entre les concepts d'œuvre et d'expression prête à confusion (Le Pape 2013).

Les bibliothèques nationales de France et d'Espagne ont complètement implémenté la structure FRBR dans leurs données LOD. La plupart des autres institutions ont publié des notices plates. Enfin, la DNB a développé, en parallèle à son modèle plat, une implémentation expérimentale du modèle BIBFRAME en LOD.

Pour les données de RERO, un modèle plat a été choisi en raison de sa simplicité de mise en œuvre. Cela n'exclut pas le passage à une structure plus hiérarchisée par la suite, selon les tendances et les besoins. Cette possibilité d'évolution future a été prise en compte dans les choix de modélisation effectués.

5.3.2 Identification des types d'entités

Afin de sélectionner les classes à attribuer aux ressources, il faut identifier les types d'entités présents. Pour ce faire, les données du catalogue collectif RERO ont été comparées avec les entités du modèle FRBR, puisque celui-ci les identifie de manière complète et structurée (voir tableau 3). La colonne *Autorités sujet RAMEAU* du tableau ne contient que les types d'entités utilisées par RERO pour son indexation.

Les données de RERO DOC n'ont pas été examinées, car leur structure ne comporte qu'un seul niveau. Elles ne sont pas liées à des notices d'autorité auteur ou sujet.

Cette analyse, bien que sommaire, permet d'effectuer plusieurs constats. Il existe divers types d'autorités, selon les étiquettes MARC21 utilisées dans le catalogue collectif RERO :

- trois types d'autorités auteur : personne, collectivité, congrès
- cinq types d'autorités sujet *noms propres* : personne, collectivité, événement, titre anonyme, nom géographique
- un type d'autorité sujet *nom commun*, dont les données sont issues de RAMEAU

Tableau 3: Types d'entités RERO

Famille FRBR		Catalogue collectif	Autorités auteur RERO	Autorités sujet RERO	Autorités sujet RAMEAU
groupe 1	item	exemplaire			
	manifestation	notice bibliogr.			
	expression				
	œuvre			titre anonyme ; (personne)	
groupe 2	personne		personne	personne	
	collectivité		collectivité	collectivité	
	famille			(personne)	
groupe 3	concept				nom commun
	objet			(personne ; titre anonyme ; événement ; nom géographique)	
	événement		congrès	événement	
	lieu		(collectivité)	nom géographique	

Premièrement, l'entité FRBR expression manque, puisqu'il s'agit du produit même de la FRBRisation. Le modèle visé pour RERO dans un premier temps étant un modèle plat, les expressions ne sont pas nécessaires.

Deuxièmement, certaines entités FRBR sont identifiées par des autorités différentes :

- Une œuvre est décrite par une autorité titre anonyme. Mais lorsqu'il s'agit d'une œuvre d'un auteur, telle que *Candide* de Voltaire, l'entrée est faite en tant qu'autorité de personne.
- Une famille est entrée en tant que personne comme autorité sujet. Il n'existe pas de famille dans les autorités auteur, car il ne s'agit pas d'une entité utilisée traditionnellement dans le catalogage AACR2. Cette entité a été introduite récemment par les FRAD.
- L'entité *objet* regroupe des organismes, structures anatomiques, objets fabriqués et substances, tels que la tour Eiffel ou Apollo 11 (IFLA 2012b, p. 21-22). Dans RERO, ces entités sont décrites au moyen d'autorités de personnes avec ajout d'une œuvre (Eiffel, Gustave. - Tour Eiffel), de titres anonymes, d'événements ou de lieux.
- Les lieux sont bien décrits comme tels dans les sujets, mais il existe également des lieux dans les autorités auteur, en tant que collectivités. Il s'agit uniquement des lieux de publication pour les livres parus avant 1800.

Enfin, plus généralement, certaines entités sont présentes à double dans RERO : en tant qu'autorités sujet et auteur. Cette situation est le reflet du catalogue sur fiches,

dans lequel on recherche soit par auteur et titre dans le fichier principal, soit par sujet dans un fichier séparé. Il est impossible de trouver, depuis un point d'accès unique, les ouvrages *par* et *sur* une même personne. L'informatique peut aider à résoudre ce genre de problèmes : il faudrait fusionner les fichiers d'autorités auteur et sujet, développés et maintenus séparément depuis leur création. Mais ceci impliquerait un important travail : développer des mécanismes comparant les chaînes de caractères et regroupant de manière automatique les entités similaires, et contrôler la qualité des alignements effectués. Dans le cadre de la publication des données en LOD, la relation entre différents types d'autorités est gérée différemment selon les bibliothèques. La BnF et la Bibliothèque royale de Suède ont par exemple utilisé une même classe (skos:Concept) pour toutes les entités sujets, qu'il s'agisse de personnes, d'œuvres ou encore de lieux. Ces entités, en tant que concepts, n'identifient pas des choses du monde réel ; ces dernières sont décrites séparément et reliées à leur *concept*. Dans la même optique, la LOC considère que ses données d'autorité sont des *noms d'entités* uniquement, et non les entités elles-mêmes ; elle les a modélisées comme telles. A l'inverse, la DNB a fusionné en avril 2012 tous ses fichiers d'autorités en prenant en compte l'émergence du web sémantique. Ainsi, le Gemeinsame Normdatei (GND) nouvellement créé décrit les ressources, aussi souvent que possible, en tant qu'objets réels plutôt que concepts. Il évite de cette manière les doublons (Svensson 2013, p. 11; Pohl 2014).

A priori, cette dernière solution semble la plus propre. Elle nécessite cependant un travail conséquent qui n'a pas pu être entrepris dans ce projet. La fusion entre ces deux types d'autorités fait néanmoins partie des objectifs actuels de RERO. Une seule classe, *skos:Concept*, a donc été retenue pour toutes les autorités sujet. Les autorités auteur ont tout de même été modélisées en tant que choses du monde réel. En résumé, les classes suivantes ont été choisies[16] :

- *dct:BibliographicResource* pour les notices bibliographiques.
- *foaf:Person* et *foaf:Organization* pour les personnes et collectivités issues des autorités auteur. Les congrès ont également reçu la classe *foaf:Organization*, car ils sont considérés comme des auteurs et non comme des événements dans le catalogage en bibliothèque[17]. *foaf:Person* et *foaf:Organization* sont des sous-classes de *foaf:Agent*.

[16] Plus d'informations sur le choix des ontologies se trouvent dans le chapitre 5.4.1.
[17] Cette vision devrait peut-être être confrontée à celle des utilisateurs afin d'en connaître la pertinence.

- *dct:Location* pour les lieux de publication issus des autorités auteur.
- *skos:Concept* pour toutes les autorités sujet.

Les items (données d'exemplaires) n'ont pas été modélisés dans ce travail, car l'intérêt de publier ce type de données n'est pas prépondérant pour le web sémantique, en tous cas dans un premier temps. En effet, il s'agit de données locales qui ne peuvent être reliées à des ressources externes. Néanmoins, leur publication est à envisager à des fins de transparence, de réutilisation des données publiques et en vue d'autres potentiels usages futurs.

5.3.3 Données de provenance

Certaines données du web sémantique peuvent être mises à disposition à plusieurs endroits différents ou sous plusieurs IRIs différents. La provenance est alors nécessaire pour déterminer quelles sont les données les plus récemment mises à jour ou les plus fiables. La thématique a fait l'objet de plusieurs groupes de travail au sein du W3C, dont le *Provenance Incubator Group* qui définit la provenance ainsi (W3C 2010, chap. 2.1) :

> *« Provenance of a resource is a record that describes entities and processes involved in producing and delivering or otherwise influencing that resource. Provenance provides a critical foundation for assessing authenticity, enabling trust, and allowing reproducibility. Provenance assertions are a form of contextual metadata and can themselves become important records with their own provenance. »*

La provenance devrait décrire d'une part le contexte de création des données (auteur, date de création, de modification, de publication, mode de transformation, etc.) et d'autre part les modes d'accès aux données (dumps, SPARQL endpoint, etc.). Plusieurs méthodes sont possibles pour ajouter ces informations lors de la publication des données (Hartig, Zhao 2010, chap. 3) :

- Créer une description du jeu de données au moyen du *Vocabulary of Interlinked Datasets* (VoID)[18].
- Ajouter la provenance aux objets LOD.
- Ajouter la provenance aux dumps (fichiers de données à télécharger).
- Fournir la provenance par le SPARQL endpoint.

Cette liste n'est pas exhaustive, d'autres pratiques existent et peuvent émerger. La thématique de la provenance n'a pas encore atteint un stade de standardisation suffisant pour pouvoir émettre des recommandations.

Il existe des vocabulaires spécifiques pour décrire les données de provenance, tels

[18] Disponible à cette adresse: http://rdfs.org/ns/void# (consulté le 14 août 2014)

que PROV-O[19], une spécification du W3C, et des vocabulaires plus généralistes, tels que les termes Dublin Core[20].

Parmi les données RDF déjà publiées par d'autres bibliothèques, les informations de provenance sous forme de descriptions VoID (pour des jeux de données entiers) se sont généralisées. A RERO, de telles descriptions ont été créées pour les divers jeux de données (exemple en annexe 3). Pour plus de précision, il a été décidé d'ajouter des informations de provenance à chaque objet LOD. Peu d'autres institutions l'ont fait. Parmi celles-ci, la BNE a utilisé le Provenance Vocabulary[21], une spécialisation de PROV-O, pour créer des données très complètes. En France, le SUDOC fournit également des informations de provenance, mais de manière sommaire au moyen de Dublin Core. Cette seconde solution a été adoptée à RERO en raison de sa simplicité et de la popularité de Dublin Core. Chaque objet LOD s'est ainsi vu attribuer un graphe, nommé *about* dans ce projet, contenant les informations de provenance suivantes :

- date de création des données de base
- date de dernière modification des données de base
- date de publication des données RDF
- créateur des données

Un exemple concret est expliqué en détail dans l'annexe 4.

5.3.4 Attribution d'IRIs

Cette étape consiste à choisir une structure logique pour les IRIs de l'ensemble des ressources RERO. Cette tâche est donc étroitement liée à l'identification des types d'entités (c.f. chapitre 5.3.2), car chacun recevra peut-être un modèle d'IRI spécifique.

> *« On the Semantic Web, URIs* [ou IRIs] *identify not just Web documents, but also real-world objects like people and cars, and even abstract ideas and non-existing things like a mythical unicorn. We call these real-world objects or things. »*
>
> (W3C 2008, chap. 3)

Au moment du design d'IRIs, il faut donc établir une distinction entre ces deux entités, car lorsque l'utilisateur déréférence un IRI identifiant une chose du monde réel, Internet ne peut pas lui transmettre cette chose physiquement, mais uniquement une représentation sous la forme d'un document web (Archer 2013, chap. 3.2.1.2).

[19] Disponible à cette adresse: http://www.w3.org/TR/2013/REC-prov-o-20130430/ (consulté le 14 août 2014)
[20] Disponible à cette adresse: http://dublincore.org/documents/2012/06/14/dcmi-terms/ (consulté le 14 août 2014)
[21] Disponible à cette adresse: http://purl.org/net/provenance/ns# (consulté le 14 août 2014)

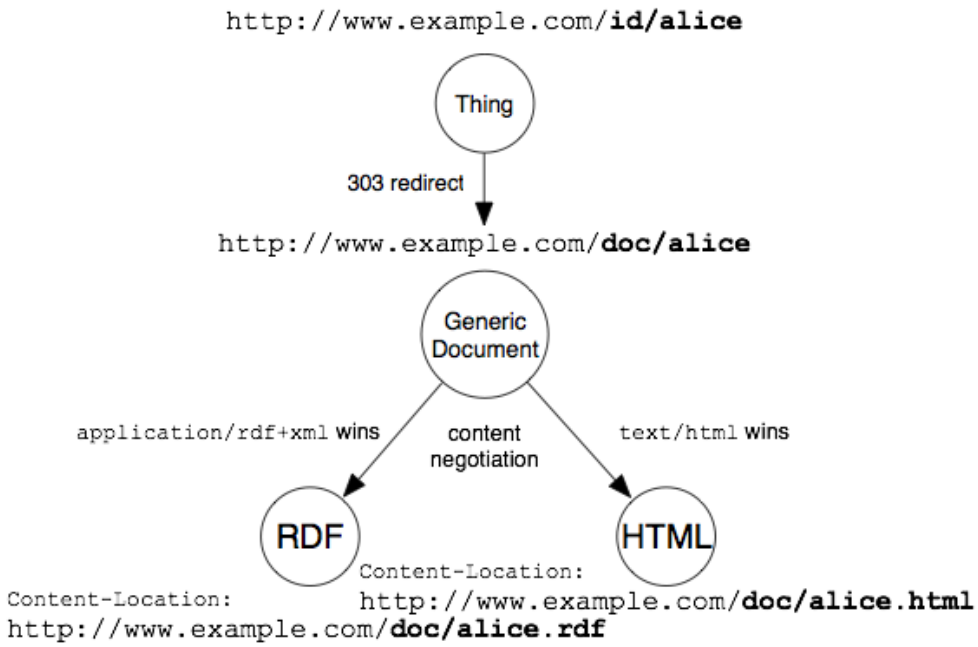
Pour le fournisseur des données, il existe deux méthodes pour mettre en place ce système (W3C 2008, chap. 4) :

- Utiliser le code HTTP 303 : si l'IRI d'une chose du monde réel est requis par un client, le serveur redirige le client vers un autre IRI, en l'occurrence celui du document web représentant l'objet.
- Utiliser un *hash-IRI* de type *www.exemple.ch/ressource#1234* pour les choses du monde réel. Lors d'une telle requête, le serveur renverra, selon le protocole HTTP, l'IRI *www.exemple.ch/ressource* tronquée de son *hash*, identifiant une représentation.

Pour les grands jeux de données, la méthode HTTP 303 est recommandée. C'est celle qui est utilisée par la plupart des bibliothèques ayant publié leurs données en RDF, et celle qui a été choisie pour RERO.

Pour une meilleure interopérabilité, le réseau a également choisi d'utiliser, tel qu'il l'est recommandé, la *négociation de contenu*. Quand un client (navigateur Web ou autre application) envoie une requête HTTP, il transmet ses préférences en terme de langue et de format (par exemple HTML, RDF/XML ou Turtle). Le serveur sélectionne alors dans le système ce qui correspond le mieux à la demande et transmet le contenu voulu. La négociation de contenu permet de fournir du HTML aux humains sur un navigateur Web et du RDF/XML aux machines (programmes informatiques), par exemple.

Figure 8: HTTP 303 et négociation de contenu



(W3C 2008, chap. 4.2)

La méthode HTTP 303 et la négociation de contenu sont schématisés dans la figure 8. L'implémentation de ces deux outils techniques nécessite le choix d'IRIs différents pour chaque étape de la requête. Pour ce faire, il n'existe à ce jour aucun standard, mais certaines institutions émettent des recommandations. Ainsi, pour les données publiques britanniques, le modèle suivant a été utilisé (Great Britain. Chief Technology Officer Council 2009, p. 9) :

http://{domain}/{doc/id}/{concept}/{reference}/{doc.file-extension}

id est utilisé pour les choses du monde réel et *doc* pour leurs représentations. *concept* correspond au type d'entité et *reference* est l'identifiant de la ressource. Enfin, l'extension du fichier, absente pour les choses du monde réel, apparaît à la fin.

Dans le cas de RERO, des URIs avaient déjà été choisis, ressemblant à celui-ci :

http://data.rero.ch/01-R219759460

Selon le modèle britannique, *data.rero.ch* est le domaine, *01* (correspondant à des notices bibliographiques) le concept et *R219759460* la référence. Il manque seulement la distinction entre chose réelle et représentation web. Celle-ci a été introduite par l'ajout, pour les représentations web, du suffixe *about*, signifiant clairement que les données fournies en décrivent d'autres. Avec l'extension de fichier, cela donne pour une notice bibliographique :

http://data.rero.ch/01-R219759460/about/rdf

La particule *01* varie en fonction du type d'entité :

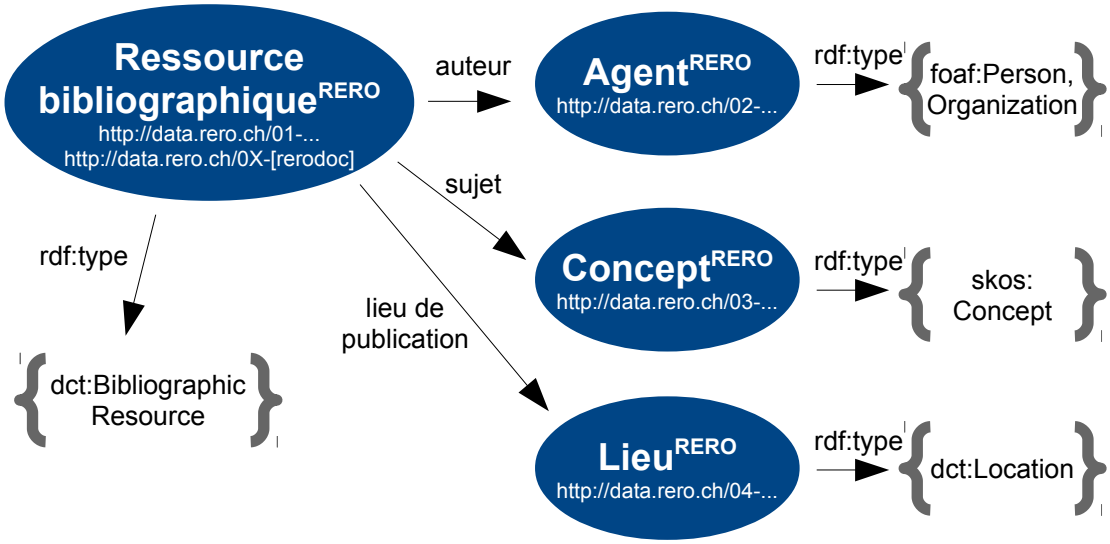
- *02* pour les auteurs
- *03* pour les concepts
- *04* pour les lieux
- Etc.

Ce modèle d'IRI, grâce à sa souplesse, pourrait très facilement être adapté à une structure des données comme FRBR, où la particule de deux chiffres indiquerait le type d'entité (œuvre, expression, manifestation, item).

5.3.5 Le modèle RERO

La modélisation s'est déroulée de manière incrémentale, tout au long du processus, en parallèle à d'autres étapes telles que l'élaboration du mapping et la transformation. Le modèle général résultant de cette phase est représenté de manière simplifiée dans la figure 9. Des représentations détaillées sont disponibles au chapitre 5.9.

Figure 9: Modèle RERO (représentation simplifiée)



5.4 Mapping

Créer un mapping entre deux formats consiste à établir des correspondances entre le format d'entrée et le format de sortie. Les correspondances sont décrites sous la forme de règles de conversion en langage humain par un spécialiste du domaine. Elles forment un mapping conceptuel qui servira de base, lors de l'étape suivante, à l'implémentation des règles en langage informatique par un programmeur. Pouvant être source de confusion, cette transition nécessite une collaboration étroite entre les deux types de participants (Geipel 2012, p. 2-4).

Pour les données de RERO, il a été choisi de créer un mapping pour chaque type d'entité identifié au chapitre 5.3.2 et illustré dans la figure 9 :

- Ressources bibliographiques (catalogue collectif)
- Agents (autorités auteur)
- Concepts (autorités sujet)
- Lieux (lieux de publication parmi les autorités auteur)

Deux mappings supplémentaires ont en outre été créés :

- Ressources bibliographiques RERO DOC, car le format d'input est légèrement différent du format MARC21 utilisé pour le catalogue collectif
- Notices *about*, car ce mapping s'applique à toutes les ressources sans distinction

Lors d'une conversion en RDF, l'établissement d'un mapping implique tout d'abord le choix des ontologies que l'on souhaite utiliser pour décrire ses données, puis la formulation de règles de conversion.

5.4.1 Choix des ontologies

Pour décrire des données RDF, des ontologies (ou vocabulaires) sont nécessaires. Deux approches sont alors envisageables (hÓra 2007) :

- L'approche microscopique : créer si possible sa propre ontologie
- L'approche macroscopique : réutiliser si possible des ontologies existantes

L'avantage de la première solution est la liberté de définir avec précision le sens des termes et leurs contraintes logiques, ainsi que de garder le contrôle sur l'évolution de l'ontologie. L'avantage de la seconde solution est d'éviter une prolifération d'ontologies sur le web qui, si elles ne sont pas reliées, créent une barrière d'interopérabilité.

Pour les données de RERO, la seconde approche a été retenue. En effet, le but de ce projet est avant tout de permettre la réutilisation des données par des tiers plutôt que d'exprimer l'exactitude des données d'origine en RDF.

Une quantité importante d'ontologies sont disponibles en ligne, y compris dans le domaine des bibliothèques. Le web sémantique n'a pas encore atteint une maturité suffisante pour une autorégulation de ces ontologies. Cette situation pouvant mener à des problèmes d'interopérabilité, des efforts d'alignement entre vocabulaires sont entrepris (Dunsire et al. 2012, p. 11). Ils consistent à établir des correspondances ou d'autres relations entre les éléments de métadonnées. Dans ce contexte, le choix de la réutilisation d'ontologies existantes peut contribuer à créer des standards de facto au sein d'une communauté comme celle des bibliothèques, qui auront ensuite plus de chance d'être reconnus par les autres communautés. A RERO, il n'est cependant pas exclu de compléter cette approche, si nécessaire, par l'élaboration d'une ontologie propre.

Vu le grand nombre d'ontologies disponibles en ligne, leur sélection est difficile. Le site web *Linked Open Vocabularies* (Vatant, Vandenbussche 2014) peut alors constituer un bon point de départ. Il répertorie et classe les vocabulaires et leurs éléments par

domaine, les rendant accessibles au moyen d'une interface de recherche. Pour identifier les ontologies du domaine des bibliothèques les plus intéressantes et les plus pertinentes pour RERO, d'autres sources[22] ont été utilisées en complément. Le tableau 4 présente brièvement les principaux vocabulaires identifiés.

Tableau 4: Ontologies pertinentes pour les bibliothèques

BIBFRAME	http://bibframe.org/vocab/ Abr. de Bibliographic Framework Transition Initiative. Ontologie en cours de développement par la LOC et permettant d'exprimer des données selon le modèle BIBFRAME (détails au chapitre 2.2.2.3). Ce modèle, dont le but est le remplacement de MARC, se veut ouvert à des données autres que celles provenant des bibliothèques. Il est par conséquent assez général.
BIBO	http://purl.org/ontology/bibo/ Abr. de Bibliographic Ontology. Ontologie publiée en 2007 et spécialisée dans la description bibliographique, notamment de livres. Elle se base sur d'autres ontologies telles que Dublin Core, et apporte des propriétés complémentaires (ISBN, édition, etc.). BIBO a été conçue et est maintenue par des particuliers.
CIDOC-CRM	http://www.cidoc-crm.org/cidoc-crm/ Abr. de Conceptual Reference Model du Comité international pour la documentation (du Conseil international des musées). Ontologie permettant d'exprimer des données muséographiques et dont le modèle est fortement axé sur le concept d'événement pour créer des liens entre les ressources. Dans ce contexte, l'ontologie FRBRoo a été créée pour rendre le modèle du CIDOC interopérable avec le modèle FRBR de la communauté des bibliothèques.
Dublin Core	Element set : http://dublincore.org/documents/dces/ Terms : http://dublincore.org/documents/dcmi-terms Ontologie la plus connue et la plus répandue sur le web. Elle fournit les métadonnées essentielles pour la description de ressources, telles que titre, auteur, date, etc. De nombreuses autres ontologies, comme BIBO, se basent sur elle et la complètent. Dublin Core est composée de deux parties : 1. L'*Element set* contient quinze propriétés génériques. Déclarées en RDF en l'an 2000 déjà, ces propriétés sont définies sans restriction de rang et de domaine. 2. Les *Terms* ont été introduits en 2006 pour compléter l'*Element set*. Ils en reprennent les quinze propriétés de base en les définissant plus précisément, et en ajoutent de nouvelles.
EDM	http://www.europeana.eu/schemas/edm/ Abr. de Europeana Data Model. Ontologie développée par la bibliothèque numérique Europeana afin de décrire des ressources numérisées provenant de divers

[22] Tableau établi sur la base des sources suivantes : (Bermès 2013; Fürste 2011, p. 82-100; Klee 2013; Library Linked Data Incubator Group (W3C) 2011, chap. 5)

fournisseurs (bibliothèques, musées, archives). EDM vise à uniformiser les métadonnées hétérogènes collectées par Europeana, tout en conservant leurs spécificités et détails.

FOAF http://xmlns.com/foaf/spec/

Abr. de Friend Of A Friend. Avec Dublin Core, l'une des ontologies les plus utilisées, axée sur la description de personnes, des relations entre elles et avec des documents. FOAF ne dépend d'aucun organisme de normalisation ; elle a été développée et est maintenue par des particuliers.

FRBR
FRBR : http://iflastandards.info/ns/fr/frbr/frbrer/frbrer/
FRAD : http://iflastandards.info/ns/fr/frad/
FRSAD : http://iflastandards.info/ns/fr/frsad/

Abr. de Functional requirements for bibliographic records. Ontologie permettant de décrire les ressources selon le modèle FRBR et ses deux compléments FRAD et FRSAD (détails du modèle au chapitre 2.2.2.1). Cette ontologie émane directement de la communauté des bibliothèques puisqu'elle a été développée par l'IFLA. Elle est donc très spécialisée sur les ressources des bibliothèques et les besoins de leurs usagers.
La version originale de l'IFLA s'étant fait attendre, plusieurs autres versions de FRBR en RDF ont été créées auparavant.

ISBD http://iflastandards.info/ns/isbd/elements/

Abr. de International Standard Bibliographic Description. Ontologie décrivant des ressources selon les règles de catalogage bibliothéconomiques ISBD élaborées par l'IFLA. Plus simple que son équivalente FRBR, l'ISBD en RDF est relativement répandue sur le web, souvent en complément des propriétés BIBO.

RDA http://www.rdaregistry.info/

Abr. de Resource Description and Access. Ontologie décrivant des ressources selon les règles de catalogage bibliothéconomiques RDA (détails au chapitre 2.2.2.2), élaborées par des associations de bibliothèques anglo-saxonnes. Les règles RDA sont basées sur le modèle FRBR et ont été conçues pour l'environnement numérique des ressources. La version RDF contient un grand nombre de propriétés, subdivisées en sept sous-ensembles d'éléments.

Schema.org http://schema.org/

Ontologie généraliste, développée par Google, Yahoo, Bing et Yandex spécialement pour que les webmasters enrichissent leurs pages avec des données sémantiques. L'avantage est que ce langage est reconnu par les principaux moteurs de recherche et permet d'en optimiser les résultats. Néanmoins, Schema.org seul ne suffit pas à une description bibliographique précise ; d'autres ontologies doivent être utilisées en complément. Pour pallier ce problème, un groupe communautaire du W3C, nommé Schema Bib Extend, travaille actuellement sur le développement d'une extension (Wallis 2013).

SKOS	http://www.w3.org/2004/02/skos/core#
Abr. de Simple Knowledge Organization System. Ontologie permettant d'exprimer en RDF de manière simple des thésaurus, classifications et autres vocabulaires contrôlés. Il s'agit d'une recommandation du W3C depuis 2009.	

Plusieurs bibliothèques ont également créé leur propre ontologie afin d'exprimer des données de manière précise selon leurs besoins. Par exemple, la bibliothèque nationale d'Allemagne a développé la *GND-Ontologie* et celle d'Angleterre le *British Library Terms RDF schema*.

Il existe en sus de nombreuses autres ontologies spécialisées dans la description bibliographique. Des ontologies centrées sur d'autres domaines (description de personnes, d'organisations, d'événements, etc.), à l'instar de FOAF, peuvent également être très utiles pour les bibliothèques.

Plusieurs éléments sont à considérer lors du choix des ontologies (Bermès 2013, chap. A.5) :

- leur auteur et la communauté à laquelle elles se rattachent
- leur niveau d'interopérabilité (dépendant de leur simplicité et de leur utilisation sur le web)
- leur précision
- leur garantie de pérennité

La popularité sur le web et l'adoption par la communauté des bibliothèques constituent, à ce stade, les principaux critères stratégiques retenus pour RERO. Une analyse comparative a été effectuée afin de déterminer les éléments et les vocabulaires les plus utilisés en bibliothèque. Ainsi, les ontologies Dublin Core et BIBO, simples et largement répandues, ont été sélectionnées. Ceci garantira un bon niveau d'interopérabilité et permettra aux données d'être comprises et utilisées par le plus grand nombre. Pour les informations plus spécifiques au domaine, le vocabulaire RDA a été préféré aux autres car il peut être utilisé avec des données qui ne sont pas forcément structurées selon FRBR[23]. De plus, une version stable de RDA a été publiée, à l'inverse de BIBFRAME par exemple, qui est encore en développement. Ce vocabulaire permet en outre de se positionner en faveur d'une évolution des modèles bibliographiques vers de nouveaux standards. D'autres ontologies spécialisées ont été utilisées ponctuellement pour certains types de données : FOAF pour les personnes,

[23] RDA fournit des propriétés dites *non contraintes*, c'est-à-dire qu'elles ne sont pas associées à des classes du modèle FRBR.

EDM pour les liens vers les documents numérisés, SKOS pour les concepts, VoID pour la description des jeux de données, etc. En général, un nombre aussi restreint que possible de vocabulaires différents a été sélectionné. Treize d'entre eux ont été retenus ; ils sont résumés dans le tableau 5.

Tableau 5: Ontologies adoptées

Ontologie	CURIE	URI
Bibliographic Ontology	bibo	http://purl.org/ontology/bibo/
DBpedia (ontologie)	dbp	http://dbpedia.org/ontology/
Dublin Core (Element Set)	dc	http://purl.org/dc/elements/1.1/
Dublin Core (Metadata terms)	dct	http://purl.org/dc/terms/
Europeana Data Model	edm	http://www.europeana.eu/schemas/edm/
Friend Of A Friend	foaf	http://xmlns.com/foaf/0.1/
RDA (unconstrained properties)	rdau	http://rdaregistry.info/Elements/u/
RDF	rdf	http://www.w3.org/1999/02/22-rdf-syntax-ns#
RDF Schema	rdfs	http://www.w3.org/2000/01/rdf-schema#
Simple Knowledge Organization System	skos	http://www.w3.org/2004/02/skos/core#
Vocabulary of Interlinked Datasets	void	http://rdfs.org/ns/void#
Web Ontology Language	owl	http://www.w3.org/2002/07/owl#

RDF et RDF Schema servent à insérer les données dans un modèle cohérent. Un CURIE, abréviation de *Compact URI*, est un moyen de raccourcir les URIs par l'utilisation de préfixes dans les données afin de les rendre plus lisibles. Le site web *prefix.cc* (Cyganiak 2014) permet de rechercher les préfixes généralement utilisés par les développeurs de RDF.

5.4.2 Règles de conversion

Les règles de conversion se présentent sous la forme de tableaux. Plusieurs bibliothèques ont mis leurs tables de conversion (ou des extraits) à disposition sur le web, comme la BnF, le SUDOC ou HBZ. Ces tables contiennent au minimum, pour chaque règle, les informations suivantes : zone MARC du format d'input et son libellé, propriété RDF correspondante, description de la règle. Bermès (2013, chap. A.6.a) propose d'insérer les quatre éléments suivants dans un mapping : source, cible, règle de mapping, exemple.

En fonction de ces recommandations, des tables ont été créées pour le mapping des données de RERO. Le tableau 6 en présente un extrait.

Tableau 6: Exemples de règles de conversion

Zone MARC source	Information transmise par le triplet	Propriété RDF	Règle de mapping	A propos de la propriété utilisée	Remarque/aide
020	ISBN	bibo:isbn10	Dans le champ $a, si les caractères formant la première chaîne suivie sont au nombre de 10, utiliser bibo:isbn10 avec cette chaîne comme objet du triplet en littéral.	Domaine: bibo:Collection ou bibo:Document. Sous-propriété de: bibo:isbn	L'ISBN est composé de 10 ou 13 caractères, pas forcément uniquement des chiffres. Le champ $a peut contenir d'autres informations telles que le type de reliure. Ex: "$a 0774803258 (pbk.)".
700	entrée secondaire (personne)	dct:contributor dc:contributor	Si un IRI de l'auteur existe, utilisation de dct:contributor avec l'IRI comme valeur. Si aucun IRI n'existe, utiliser dc:contributor avec comme valeur la concaténation des valeurs de $a $b $c $d $q (ignorer tout ce qui apparaît dans l'ordre après $t). Avant les sous-champs $b $c $d $q, s'ils existent, insérer un espace. Répéter ces règles pour toutes les zones 700.	dct:contributor: "An entity responsible for making contributions to the resource." dct:creator: "An entity primarily responsible for making the resource". Sous-propriété de dct:contributor et de dc:creator. Rang: dct:Agent.	Toutes les entrées de type "auteur" (100 et 700) se font à dc:contributor, car aucun code de rôle n'est utilisé dans les règles RERO. Il est donc impossible de distinguer les auteurs principaux (dc:creator) des autres (dc:contributor) tels que les éditeurs, les traducteurs, etc. La propriété "dc:contributor" a été préférée à "dc:creator", car la première a pour sous-classe la seconde. Cela signifie que les "creator" sont également des "contributor". Alternatives envisagées: utiliser dc:creator pour toutes les zones, ou pour les zones 100 seulement.

Deux colonnes y ont été ajoutées :

- Des informations sur la propriété concernée (rang, domaine, classe supérieure, etc.) afin d'éviter des incohérences dans le modèle.
- Des remarques servant généralement à justifier les choix effectués en prévision d'une lecture ou d'une révision du mapping par d'autres personnes. Ceci permet également à la personne qui implémente les règles en langage informatique de comprendre certaines particularités des données d'origine.

La deuxième règle du tableau 6 donne un exemple concret de ces situations. Les règles de catalogage en vigueur à RERO ont subi des simplifications excessives par le passé. De ce fait, aucun moyen ne permet actuellement de distinguer les auteurs principaux des auteurs secondaires (éditeurs, traducteurs, etc.), ce qui est possible dans certaines autres bibliothèques, où des codes de rôles sont utilisés. Néanmoins, la zone 100 (entrée principale pour personne) contient toujours un auteur principal si la ressource en possède un. Faut-il par conséquent utiliser une propriété RDF différente pour les personnes en entrée principale et en entrée secondaire ? Après débat, analyse et réflexion, la décision a été prise de n'utiliser qu'une seule propriété afin

d'éviter de créer des données peu logiques. Le choix s'est porté sur la propriété *dct:contributor* lorsqu'une autorité existe pour l'auteur concerné, signifiant qu'il possédera un IRI RERO, et sur *dc:contributor* lorsqu'aucune autorité n'existe. En effet, *dct:contributor* a pour rang *dct:Agent* et ne peut donc avoir un littéral comme objet, à l'inverse de *dc:contributor*.

Comme le montre la quatrième colonne du tableau 6, les règles de conversion peuvent contenir des opérations d'analyse et de traitement des données : analyse de la zone 020 afin de ne prendre en compte que les ISBN à 10 chiffres, suppression ou ajout de caractères (souvent de la ponctuation), concaténation de sous-champs, etc. L'exemple montre que les règles peuvent également être conditionnelles.

Les six mappings de RERO ont été développés progressivement et parallèlement. Les règles ont été formulées d'abord pour les données considérées comme prioritaires, selon leur niveau de contrôle et la pertinence de leur contenu pour l'identification d'une ressource. Une table de conversion est à comprendre comme un document de travail, adaptable selon les besoins. Pour le projet de RERO, elles ont été complétées par d'autres informations au sein de colonnes additionnelles, à des fins de communication entre les participants et de gestion du processus, notamment pour les étapes de la transformation et du contrôle qualité.

5.5 Liens externes

La notion de *liens externes* correspond, dans ce travail, aux relations entre des ressources RERO et des ressources externes au réseau. La création de ces liens s'effectue en général de deux manières différentes.

La première est la génération de liens directement d'après les données de base, au moyen d'un identifiant qui y est stocké. Cette méthode est simple à mettre en œuvre, mais le choix des jeux de données à relier est restreint. Les liens produits ne sont donc pas toujours les plus intéressants, mais ils sont très fiables, si les identifiants utilisés dans les données de base sont soumis à un contrôle strict au sein du système.

La seconde méthode est la génération de liens par alignement de deux jeux de données, en comparant les chaînes de caractères des entités que l'on souhaite relier. L'alignement peut être réalisé à l'interne par le propriétaire des données, ou à l'externe par un prestataire de services comme VIAF (c.f. chapitre 5.5.1). Il se fait si possible de manière automatisée pour les grand jeux de données, au moyen de mécanismes développés sur mesure ou de logiciels spécialisés tels que Silk (Isele et al. 2014). Un

alignement manuel – effectué par des personnes – est envisageable dans le cas où une liste interne restreinte de valeurs contrôlées est utilisée. Enfin, il est également possible d'effectuer des alignements semi-automatisés : certains logiciels comme OpenRefine permettent de réconcilier deux jeux de données, mais les résultats doivent ensuite être validés.

Généralement, cette seconde méthode est moins précise que la première et peut mener à des erreurs, mais elle peut produire des liens plus intéressants, car les jeux de données externes à relier sont choisis.

Dans le cas de RERO, les liens ont été générés directement durant la transformation, selon les règles de conversion, et principalement au moyen de la première méthode. Les données ont été liées à plusieurs référentiels du web sémantique.

5.5.1 VIAF

Les données sur les auteurs et les collectivités ont été reliées aux données du VIAF (Virtual International Authority File). Ce portail, maintenu par OCLC, fournit un accès centralisé aux notices d'autorité auteur de différents réseaux et grandes bibliothèques, reliées entre elles (OCLC 2014b). RERO en fait partie. Les liens sont donc établis régulièrement par les algorithmes du VIAF et ont pu être récupérés pour ce projet. Une fois les notices RERO reliées à celle du VIAF, il devient très facile de générer des liens vers tous les autres participants au projet, parmi lesquels les bibliothèques nationales de Suisse, d'Allemagne, de France, des Etats-Unis, du Royaume-Uni, ainsi que l'encyclopédie collaborative Wikipédia en anglais. Ces jeux de données, et en particulier le dernier mentionné, sont des référentiels de données RDF internationalement reconnus et adoptés. Ils constituent une porte d'entrée très intéressante vers le web sémantique.

5.5.2 RAMEAU

RAMEAU, ou Répertoire d'autorité-matière encyclopédique et alphabétique unifié, contient les notices d'autorité sujet de la BnF (2014c). Celles-ci possèdent de nombreuses correspondances avec des notices équivalentes du GND allemand et du LCSH américain[24].

RERO utilise les noms communs de RAMEAU depuis 2011 pour son indexation. De nombreux liens ont donc été créés avec ce référentiel lors de la saisie des données par les catalogueurs du réseau. Leur qualité est donc bonne. RAMEAU intègre très bien

[24] Il s'agit des fichiers d'autorités – équivalents à RAMEAU – entretenus par la DNB et la LOC.

les données de RERO dans le web sémantique, offrant des possibilités de liens vers des termes équivalents en allemand et en anglais.

5.5.3 GeoNames

GeoNames est une base de données géographique disponible en LOD. Pour les lieux qu'elle référence, elle contient, entre autres, des noms en plusieurs langues, l'altitude, le nombre d'habitants, la classification politique, le numéro postal ou les coordonnées géographiques (*About GeoNames* 2014). GeoNames constitue l'un des nœuds les plus importants du web sémantique (Jentzsch, Cyganiak, Bizer 2011).

La mention du pays de publication d'un document est disponible dans les données de base sous forme de codes de pays MARC, dont la liste est maintenue par la LOC. Ce référentiel n'étant que peu reconnu hors de la communauté des bibliothèques, un mapping a été réalisé avec les identifiants GeoNames. Il a pu être créé de manière automatique grâce à une table de correspondances entre les codes de pays MARC et ISO 3166-2, mise à disposition par l'Open Knowledge Foundation.

De plus, le catalogue collectif RERO possède, dans les notices de documents édités en Suisse, un code de deux lettres identifiant le canton de publication. Des correspondances GeoNames ont été attribuées manuellement à ces codes. Ces liens valorisent ainsi les notices de documents suisses et contribuent à faire de RERO une référence de qualité pour leur mise à disposition.

5.5.4 Classification décimale universelle

La CDU (Classification décimale universelle) est l'une des classifications génériques les plus utilisées dans le monde, avec la Dewey. Elle est multilingue et un extrait de 2'600 classes est publié en LOD (UDC Consortium 2014).

Il a été envisagé de relier les données de RERO à la classification Dewey, plus connue et plus utilisée dans le web sémantique. Un champ lui est d'ailleurs dédié dans les notices MARC (numéro 082). L'indice de classification n'est toutefois pas saisi systématiquement lors du catalogage des ressources dans RERO. Environ 0,82 % des notices le possèdent tout de même, mais il s'agit de notices importées, et la qualité tout comme la syntaxe de la saisie varient trop pour pouvoir établir des liens fiables. Le champ 080, comportant quant à lui l'indice CDU, n'est présent que dans 0,11 % des notices du catalogue collectif. Néanmoins, il est obligatoire dans toutes les notices RERO DOC et son utilisation est contrôlée par une liste restreinte d'indices autorisés. Une règle de conversion a donc été ajoutée au mapping de RERO DOC, indiquant la

procédure de création de lien :

- Téléchargement du dump de la CDU
- Comparaison de la zone MARC21 080, sous-champ $a, avec les propriétés skos:notation du dump
- Si une équivalence exacte est constatée, récupération de l'IRI du concept CDU
- Si aucune équivalence n'est trouvée, pas de création de lien

5.5.5 Lexvo.org

Lexvo.org est un référentiel du web sémantique comportant des descriptions détaillées sur les langues et les éléments liés à ce sujet. Pour chaque description d'une langue, les données peuvent contenir entre autres le nom en de nombreuses langues, des liens vers des sites spécialisés, des relations avec d'autres langues, et les codes selon les diverses normes ISO (639-2, 639-3, 639-5) (Melo 2014).

Dans RERO, les données utilisées proviennent d'une liste de codes spécifique à MARC21, quasi identique aux codes ISO 639-2. Ces deux listes sont gérées par la LOC. Grâce à la mise à disposition par Lexvo.org d'un fichier de correspondances entre les identifiants Lexvo.org et les codes ISO 639-2, un mapping des codes utilisés dans RERO a pu être effectué avec le référentiel.

5.5.6 Référentiels RDA

Les référentiels RDA ont été utilisés ponctuellement pour compléter les descriptions des documents. Les classes Dublin Core ont servi de référence pour les types de contenus (texte, image, son...), et les classes BIBO pour les types de documents (monographie, périodique...). Mais dans certains cas comme la musique imprimée et manuscrite, ainsi que les médias de type *microforme*, aucune classe adéquate n'a été trouvée. En complément, les référentiels RDA, pointus pour ce genre d'aspect, ont permis de décrire ces documents en toute précision.

L'adoption de tous les types de contenus, de médias et de supports matériels RDA pour les ressources RERO, n'a pas été retenue dans ce projet. Cela aurait nécessité une analyse approfondie des données de base, voire un traitement avant la conversion, car il n'existe pas toujours une équivalence de un à un entre les référentiels RDA et les codes utilisés dans les données de RERO. Cette tâche est toutefois prévue dans un deuxième temps, lorsque le réseau pratiquera les règles RDA.

Puisque Dublin Core et BIBO présentent l'avantage d'être plus largement reconnus sur

le web, il sera envisageable de les utiliser pour les acteurs externes au monde des bibliothèques, en parallèle aux référentiels RDA pour les acteurs spécialisés souhaitant des données très précises.

5.6 Transformation

La transformation est l'étape durant laquelle les règles de conversion en langage humain sont transférées en langage informatique. Les notices et les relations entre elles sont extraites du système de gestion de bibliothèque et converties en RDF. Cette étape a été réalisée par le spécialiste de RERO. Elle ne représente pas le centre de cet ouvrage et n'y sera donc pas décrite en détail.

Comme les autres tâches, la transformation s'est faite peu à peu, en parallèle à la création des mappings conceptuels. Pour que ces deux activités se coordonnent au mieux, les règles de conversion ont été complétées par les informations suivantes : règle achevée (oui/non), date d'actualisation de la règle, date d'implémentation de la règle. Grâce à ces données, un suivi a pu être effectué sur le processus, en particulier afin d'être sûr que chaque règle est implémentée selon sa dernière actualisation.

5.7 Contrôle qualité

Au cours de la génération des premières données RDF, un contrôle de leur qualité a été effectué. Ceci a permis notamment de détecter de petites erreurs de caractères : un espace ou une barre oblique en trop par exemple.

Pour rendre cette procédure systématique, une colonne a été ajoutée dans les divers mappings, contenant un exemple de notice illustrant le mieux possible chaque règle de conversion. Pour certaines zones complexes, telles que la zone contenant le titre et la mention de responsabilité, plusieurs exemples sont nécessaires. Ainsi le contrôle qualité peut s'effectuer par règle de manière plus rapide.

Afin de ne pas effectuer de contrôle sur la base d'un échantillon de trop petite taille (une à dix notices), il est envisageable de générer pour chaque règle, en complément, une table contenant les données de base et les données RDF générées, et ceci pour une centaines de notices.

5.8 Publication des données

En fin du cycle de création des données RDF intervient la publication. L'une des principales plus-values des Linked Data est qu'elles peuvent être traitées automatiquement par des machines. Ces dernières peuvent accéder aux données par

diverses méthodes, dont les plus courantes sont (W3C 2014b, chap. 8) :

- Le déréférencement d'IRIs
  Les données de base sont transformées en RDF au moment de la requête et les mécanismes de négociation de contenu livrent alors la sérialisation demandée par le client.
- Une API[25]
  Elle permet à des clients d'effectuer certaines recherches dans le jeu de données, selon des critères prédéfinis.
- Un SPARQL endpoint
  Il permettra à des clients externes d'effectuer des recherches précises dans le jeu de données.
- Un ou plusieurs dumps à télécharger
- Les données intégrées dans les pages HTML (au moyen de la syntaxe RDFa par exemple), notamment pour les moteurs de recherche.

Pour améliorer la visibilité auprès des machines, il est également recommandé d'annoncer le jeu de données nouvellement publié aux principaux moteurs de recherche sémantiques (W3C 2013e, chap. 1.8.8.2).

L'annonce doit également s'adresser aux personnes, grâce à des messages destinés au public habituel de l'institution, à la communauté des bibliothèques et aux développeurs RDF. Pour une plus grande visibilité, le jeu de données peut être mis à disposition sur des plates-formes prévues à cet effet, comme le *Datahub*[26] ou des portails nationaux ou régionaux d'Open Government Data. Enfin, il est aussi possible de faire apparaître les Linked Data sur une interface préexistante destinée aux humains, par exemple grâce à l'affichage d'un lien RDF ou de possibilités d'export auprès de chaque notice.

A RERO, le projet a subi des retards dus à des causes qui lui sont externes. L'étape de la publication n'a donc pas encore été effectuée.

[25] Abr. de Application Programming Interface (Interface de programmation des applications). Ensemble de commandes, fonctions et protocoles proposés à travers une interface via laquelle un logiciel peut offrir des services à d'autres logiciels.

[26] Disponible à cette adresse: http://datahub.io/ (consulté le 7 août 2014)

5.9 Résultats intermédiaires

La figure 10 (pour les ressources bibliographiques) et la figure 11 (pour les autorités) représentent, de manière visuelle, le modèle de données RERO développé à ce stade du projet, accompagné des propriétés et classes utilisées. Les différents éléments sont identifiés de la manière suivante :

- Les ovales bleus sont des ressources RERO, c'est-à-dire des ressources possédant un IRI du domaine du réseau.
- Les ovales violets sont des ressources externes, provenant de référentiels du web sémantique.
- Les flèches correspondent aux propriétés des diverses ontologies retenues. Les abréviations (CURIE) de ces dernières sont développées dans le tableau 5.
- Les termes entourés d'accolades sont les classes du modèle, également issues des ontologies.
- Les rectangles gris représentent des littéraux.

Selon les mêmes conventions visuelles, la figure 12 illustre la modélisation des données de provenance. Dans cette figure, les instances RERO sont définies (*rdfs:isDefinedBy*) par des notices *about*. Ces dernières ont pour sujet (*foaf:primaryTopic*) soit des instances, soit des jeux de données. Les notices *about* d'instances font partie (*void:inDataset*) des divers jeux de données de RERO.

Figure 10: Modèle RERO : ressources bibliographiques

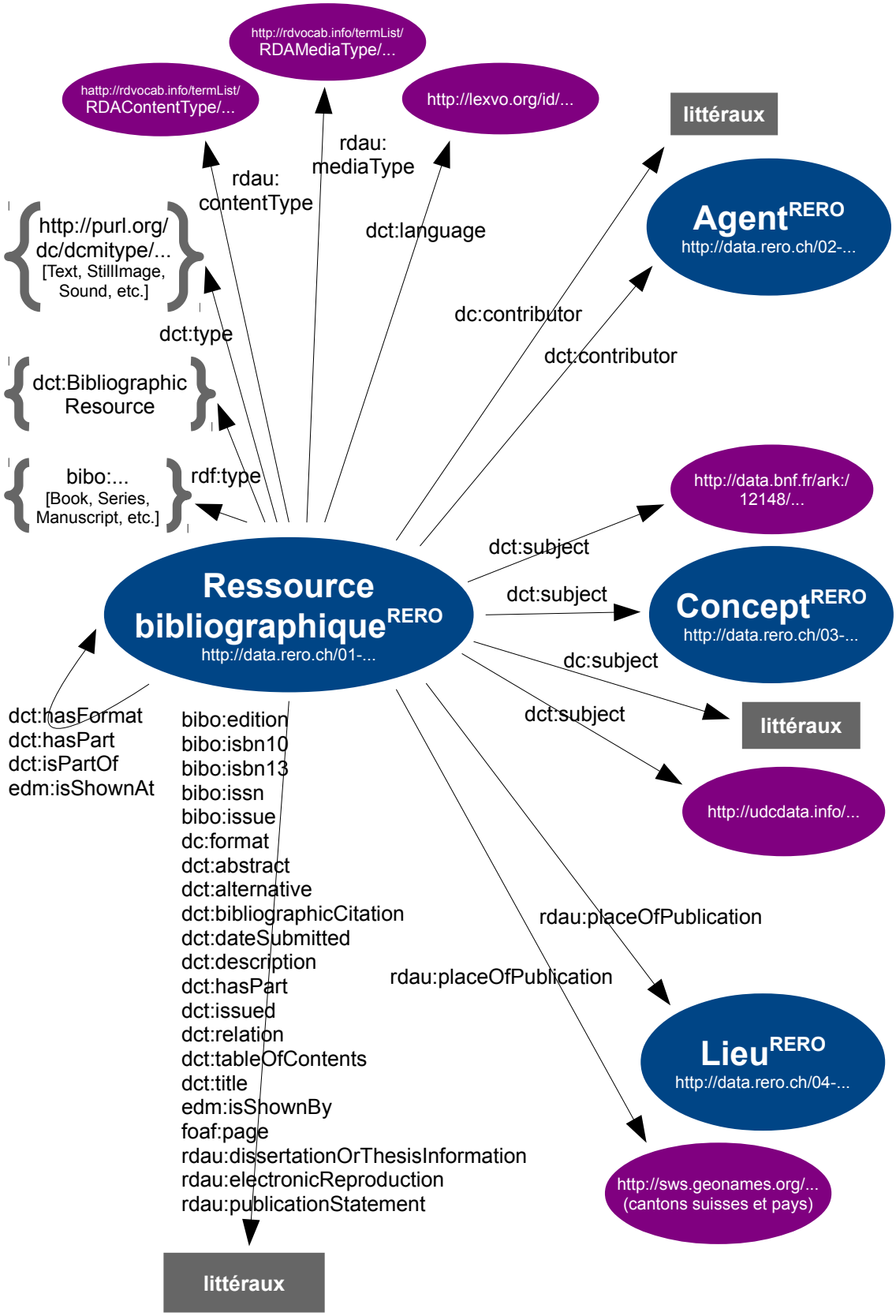


Figure 11: Modèle RERO : autorités

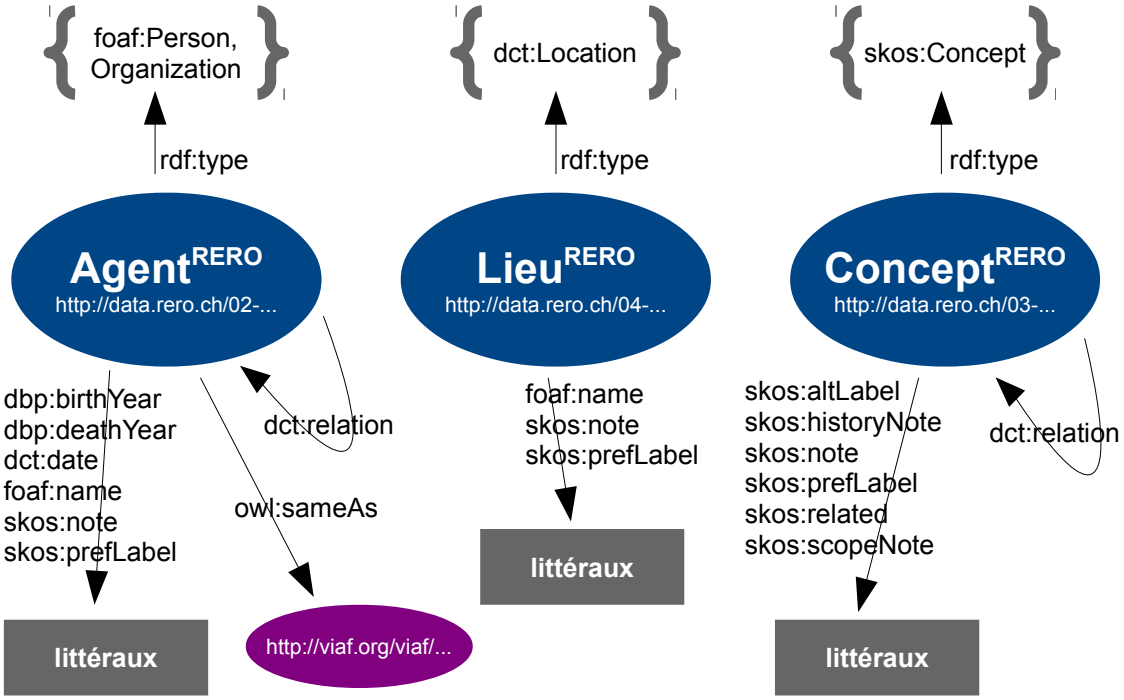
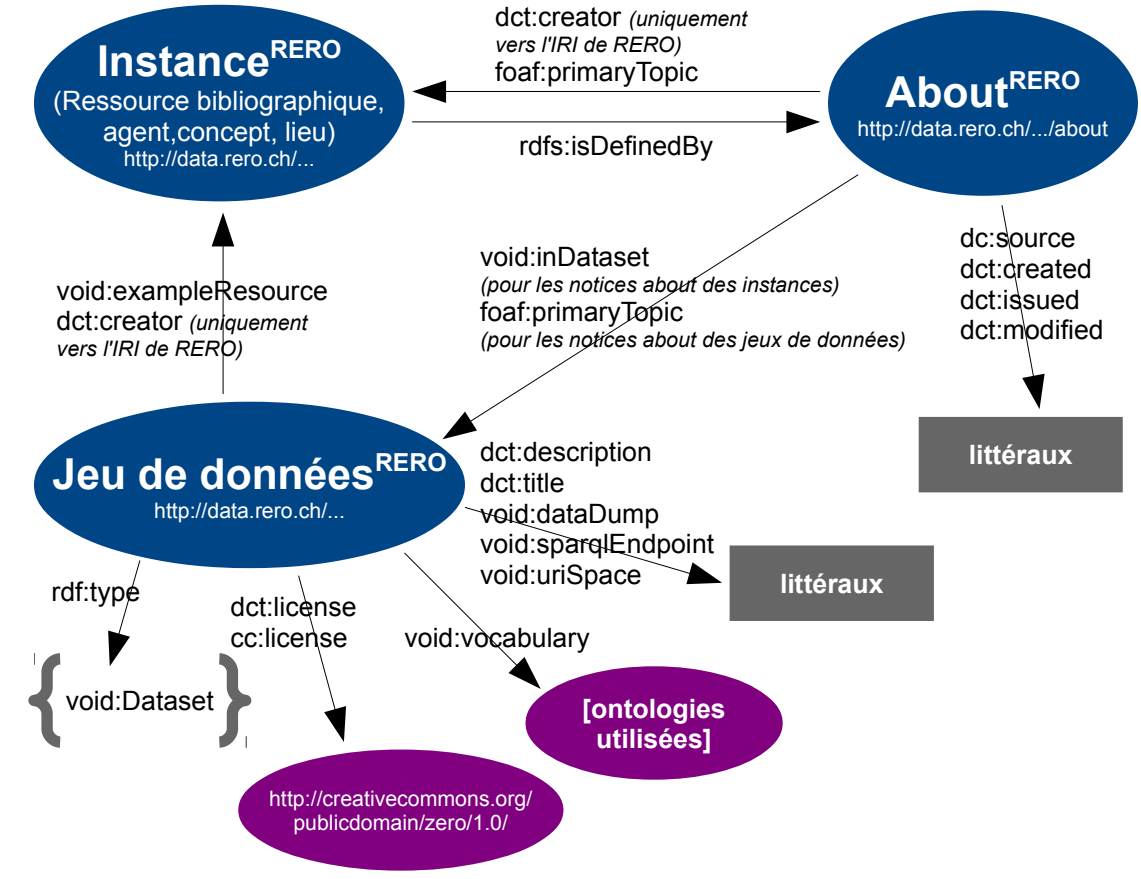


Figure 12: Modèle RERO : données de provenance



A titre d'exemple pour ce travail, une notice RERO en format MARC21 (figure 13) a été convertie en RDF. Dans ces données de base, les informations les plus importantes apparaissent de cette manière : auteur principal en zone 100, titre et mention de responsabilité en 245, adresse bibliographique en 260, collation en 300, mention de collection en 490, sujets dans les zones 6XX, auteur secondaire en 700 et lien avec la notice de collection en 830. Ce genre de notice est typique du catalogue collectif RERO.

Figure 13: Transformation : notice de base MARC21

LDR			01230nam a2200337 a 4500
001			vtls007689758
003			RERO
005			20140717143200.0
008			140326s2014 fr 00 fre d
020			$a 9782755507256
035			$a R007689758
039		7	$b 6086
039		9	$a 201407171432 $b 6600 $c 201407171424 $d 6600 $c 201405021223 $d 6086 $c 201404290859 $d 6097 $y 201403261347 $z 6042
040			$a RERO gevbge
072		7	$a s1ss $2 rero
100	0		$a Voltaire
245	1	0	$a Pensées végétariennes : $b [recueil inédit] / $c Voltaire ; éd. établie, notes et postface par Renan Larue
260			$a [Paris] : $b Milles et une nuits, $c 2014
300			$a 69 p. : $b ill. ; $c 15 cm
490	1		$a La petite collection $v 632
600	0	7	$a Voltaire $2 rero
650		7	$a Végétarisme $2 rero
700	1		$a Larue, Renan
830		0	$a Mille et une nuits. $p Ed. Mille et une nuits $v 632
902			$a gevbge $b 2014/05 $c lar040
957			$a gevbge gevimv
982			$2 ge-vimv $a BA 2014/1
992			$a BGE Bsm 5123 $x ge/vbge/a/2014/321499.5
992			$a BA 2014/1 $x ge/vimv/d/2014/2710

La figure 14 représente cette même notice en RDF. Le premier bloc, introduit par *rdf:RDF*, déclare les vocabulaires utilisés et leurs CURIE. Le second bloc fournit les données de provenance de la notice. Enfin, le dernier bloc (le plus grand) contient la notice elle-même. Les littéraux apparaissent en noir, tandis que les IRIs sont en violet.

Cet exemple de notice, tout comme les exemples se trouvant en annexes, sont exprimés au moyen de la sérialisation RDF/XML. Ce format, bien qu'il ne permette pas à l'œil humain de distinguer à première vue les triplets, est le plus répandu sur le web. Il peut être transformé de manière automatique en différentes sérialisations RDF grâce, entre autres, à des convertisseurs gratuits disponibles en ligne.

Figure 14: Transformation : notice convertie en RDF/XML

```
<?xml version="1.0"?>

<rdf:RDF
        xmlns:bibo="http://purl.org/ontology/bibo/"
        xmlns:dc="http://purl.org/dc/elements/1.1/"
        xmlns:dct="http://purl.org/dc/terms/"
        xmlns:foaf="http://xmlns.com/foaf/0.1/"
        xmlns:rdau="http://rdaregistry.info/Elements/u/"
        xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
        xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
        xmlns:skos="http://www.w3.org/2004/02/skos/core#"
        xmlns:void="http://rdfs.org/ns/void#"
        >

        <rdf:Description rdf:about="http://data.rero.ch/01-R219759460/about">
                <foaf:primaryTopic rdf:resource="http://data.rero.ch/01-R219759460"/>
                <dct:creator rdf:resource="http://data.rero.ch/02-A005399379"/>
                <dct:created
rdf:datatype="http://www.w3.org/2001/XMLSchema#dateTime">2014-03-
26T13:47:00+01:00</dct:created>
                <dct:modified
rdf:datatype="http://www.w3.org/2001/XMLSchema#dateTime">2014-07-
17T14:32:00+01:00</dct:modified>
                <dct:issued
rdf:datatype="http://www.w3.org/2001/XMLSchema#dateTime">2014-08-
31T23:59:00+01:00</dct:issued>
                <void:inDataset
rdf:resource="http://data.rero.ch/catalogue_collectif"/>
                <rdf:type rdf:resource="http://purl.org/ontology/bibo/Document"/>
        </rdf:Description>

        <rdf:Description rdf:about="http://data.rero.ch/01-R219759460">
                <rdf:type rdf:resource="http://purl.org/dc/terms/BibliographicResource"
/>
                <rdf:type rdf:resource="http://purl.org/ontology/bibo/Book" />
                <dct:type rdf:resource="http://purl.org/dc/dcmitype/Text" />
                <dct:contributor rdf:resource="http://data.rero.ch/02-A000173676" />
                <dct:title>Pensées végétariennes : [recueil inédit] / Voltaire ; éd.
établie, notes et postface par Renan Larue</dct:title>
                <rdau:publicationStatement>[Paris] : Milles et une nuits,
2014</rdau:publicationStatement>
                <rdau:placeOfPublication
rdf:resource="http://sws.geonames.org/3017382/" />
                <dct:issued>2014</dct:issued>
                <dc:format>69 p. : ill. ; 15 cm</dc:format>
                <dct:bibliographicCitation>La petite collection ;
632</dct:bibliographicCitation>
                <dct:subject
rdf:resource="http://data.bnf.fr/ark:/12148/cb11928669t" />
                <dct:subject
rdf:resource="http://data.bnf.fr/ark:/12148/cb11933774p" />
                <dc:contributor>Larue, Renan</dc:contributor>
                <dct:isPartOf rdf:resource="http://data.rero.ch/01-1851619" />
                <bibo:isbn10>9782755507256</bibo:isbn10>
                <dct:language rdf:resource="http://lexvo.org/id/iso639-3/fra" />
                <foaf:page rdf:resource="http://data.rero.ch/01-
R219759460/about/html" />
                <rdfs:isDefinedBy rdf:resource="http://data.rero.ch/01-
R219759460/about" />
        </rdf:Description>

</rdf:RDF>
```

6. Discussion et perspectives

6.1 Le perfectionnement et la pérennisation du service

Le processus de publication des données en RDF suit un cycle continu qui se poursuit également après l'étape de la publication. Ceci sert à maintenir et à pérenniser le service, afin qu'il reste moderne et de qualité, et qu'il puisse faire office de référence au sein du web sémantique. Dans cette optique, les activités suivantes sont à entreprendre régulièrement :

- Se tenir constamment informé des nouveautés, en particulier concernant les normes, les technologies et les usages.
- Adapter la modélisation des données RDF en fonction des nouveaux standards émergents.
- Réviser les ontologies utilisées, qui peuvent être modifiées, devenir obsolètes ou être abandonnées par la communauté.
- Retoucher les règles de conversion en fonction du choix des standards de catalogage utilisés dans le réseau. Eventuellement étendre les règles à de nouvelles données qui n'avaient pas été converties au début du projet, par exemple les données locales.
- Ajouter des liens externes plus pertinents, selon l'évolution du contenu et de la popularité des référentiels du web sémantique. Dans ce but, envisager la mise en place de mécanismes d'alignement des données.

Un tel service implique une charge de travail supplémentaire pour l'organisation qui le propose. Il doit être pleinement intégré à son fonctionnement, par exemple en ajoutant les activités qui en découlent dans les cahiers des tâches du personnel et en planifiant sur un ou deux ans leur réalisation. La pérennisation du service nécessite aussi la mise à jour des données fournies.

6.1.1 Mise à jour des données

Si l'on accède aux données par déréférencement, elles sont converties en RDF au moment de la requête. Dans les autres cas (SPARQL endpoint, dump, etc.), des mises à jour régulières sont nécessaires. La mise à jour des données représente un processus à part entière, constitué des différentes opérations de l'étape de transformation (c.f. chapitre 5.6). En effet, les données de base en format MARC21 sont actualisées quotidiennement par les catalogueurs du réseau, au contraire des données RDF. Ce processus doit donc être effectué le plus régulièrement possible pour la qualité du service. Il doit en outre prendre en compte et gérer les modifications dans les données de base qui peuvent avoir un impact sur les identifiants pérennes

publiés dans les données RDF, par exemple les suppressions de notices.

L'actualisation des données dans les autres bibliothèques proposant des services LOD se fait selon diverses périodicités, allant de quotidiennement à mensuellement. Le processus de mise à jour doit donc être automatisé autant que possible et devenir une opération de routine à RERO. Pour la cohérence des liens internes, il doit s'appliquer à chaque fois à tous les ensembles de données.

Travailler avec deux formats en parallèle – MARC21 pour l'enregistrement et la gestion des données, et des sérialisations RDF pour leur publication – mène à devoir effectuer des conversions récurrentes et chronophages. Quelle que soit l'institution concernée, cette situation n'est idéale ni souhaitable à long terme. Le réseau Libris en Suède a fait le pari risqué d'abandonner le format historique des données de bibliothèques en faveur de RDF, suivant la maxime « Linked Data or die » (Malmsten 2013, p. 4). En 2008, il avait déjà été le premier acteur à publier l'ensemble d'un catalogue de bibliothèque en LOD. A présent, le réseau est en phase de développement d'une interface de catalogage basée directement sur les données RDF.

6.1.2 Et après ?

Sur la base des LOD, de nombreuses fonctionnalités à plus-value pour l'utilisateur peuvent être développées (Hügi, Prongué 2013) : recherches sur une carte géographique, recherches multilingues, recherches fédérées, enrichissement des notices du catalogue, etc.

Pour les implémenter à RERO, deux voies sont à considérer.

La première est l'intégration de ces nouvelles fonctionnalités dans l'interface existante. Il faut pour cela insérer certaines données RDF, notamment les liens externes, dans les notices MARC afin qu'elles puissent être exploitées par l'interface publique de recherche, ou alors développer un mécanisme faisant le lien entre les notices MARC et les enrichissements RDF disponibles. Les liens VIAF par exemple donnent souvent accès aux articles Wikipédia correspondants, qui peuvent contenir des informations intéressantes à ajouter au catalogue en ligne.

La seconde voie est le développement d'une application indépendante de l'interface existante, basée sur les données RDF. Cette solution nécessite toutefois d'importants moyens. Seules quelques grands projets de ce type existent à ce jour, tels que *data.bnf.fr* en France ou *Kulttuurisampo* en Finlande.

Ces deux voies impliquent un investissement considérable de la part de RERO. A

moyen terme, la situation idéale serait l'utilisation d'un système de gestion de bibliothèque prenant en charge les données indépendamment de leur format, tel que le logiciel développé en Suède. L'émancipation de ces systèmes par rapport au format MARC devrait constituer un critère de sélection important pour la gestion future des métadonnées en bibliothèque.

L'effort de transparence démontré par ce projet pourrait également permettre la mise à disposition de données jusqu'à présent inaccessibles aux utilisateurs, telles que les notices d'autorité auteur et sujet. Celles-ci contiennent parfois des informations régionales rares pouvant intéresser un public ne fréquentant pas forcément les bibliothèques. Par exemple, les dates de vie de personnalités régionales sont des données factuelles présentes dans les autorités auteur. Une interface simple mettant à disposition ces données en RDF et en HTML, assortie d'un moteur de recherche pour les consulter, représenterait un bénéfice intéressant pour l'usager et permettrait de valoriser sur le web le travail des bibliothèques. De tels systèmes sont entre autres proposés par la BnF, le SUDOC et la LOC.

En général, la publication de données en LOD est bénéfique pour les utilisateurs, mais aussi pour l'institution responsable.

> *« The KB* [Bibliothèque royale des Pays-Bas] *is looking to obtain return on investment in two areas - enhanced discoverability and data enrichments - which will help the Library to remain relevant in a web-centric world. »*
>
> *(The European Library 2014a)*

L'aspect découverte (*discoverability*) du retour sur investissement implique également une amélioration de la visibilité, grâce à un meilleur référencement par les moteurs de recherche, aux IRIs permettant à des liens tiers de pointer plus facilement vers le catalogue et à la réutilisation gratuite des données.

6.2 Vers une évolution des données de RERO

Les diverses étapes du projet ont mis en évidence certaines caractéristiques problématiques des données de base, issues bien souvent des règles de catalogage utilisées.

Ainsi, la distinction entre les données et leur présentation n'est pas toujours claire. Pour séparer les sous-champs par exemple, les règles demandent aux catalogueurs, en plus des codes de sous-champs, d'insérer des signes de ponctuation spécifiques. Lors d'une conversion en RDF, ces signes, faisant partie du contenu des sous-champs, doivent être supprimés. La zone du titre (numéro 245), l'une des plus importantes et

des plus complexes du format, est particulièrement concernée par cette situation. De tradition, elle retranscrit fidèlement la page de titre et la mention de responsabilité telles qu'elles apparaissent sur le document. Par conséquent, certaines informations saisies sont difficilement traitables automatiquement. Le cas d'un ouvrage avec plusieurs titres ou avec des titres parallèles illustre bien ce fait : selon les situations, certains titres peuvent être saisis dans le sous-champ de la mention de responsabilité, en toute conformité avec les règles de catalogage AACR2. Un autre cas problématique consiste en la méthode de liens entre notices, basée sur des chaînes de caractères. Lors d'un lien entre un article et sa revue, le premier doit posséder, dans une zone spécifique, le titre exact de la revue tel qu'il apparaît dans la notice de la revue. Si le titre de cette notice fait l'objet d'une correction, ne serait-ce que d'un caractère, le lien est brisé.

La plupart de ces problèmes ne peuvent être résolus facilement car ils dépendent de règles de catalogage utilisées dans des bibliothèques du monde entier. Modifier ces règles met non seulement RERO en contradiction avec les standards de la communauté, mais crée aussi des incohérences avec les données saisies selon les anciennes règles au sein du réseau. Ces données ne peuvent généralement pas être converties de manière automatique. De tels changements sont donc à effectuer avec précaution.

Néanmoins certaines actions ciblées sur le logiciel et la gestion des données pourraient être entreprises afin d'améliorer la qualité des notices :

- Supprimer les signes de ponctuation entre les sous-champs lors de la saisie. Configurer le logiciel de gestion de bibliothèque afin qu'il les ajoute automatiquement lors de l'affichage.
- Utiliser plus de champs contrôlés grâce à des identifiants internes, notamment pour les liens entre notices.
- Utiliser des codes de rôles pour les personnes signalées dans les notices, afin entre autres de pouvoir distinguer les auteurs secondaires des auteurs principaux (c.f. problématique présentée dans le chapitre 5.4.2).

Pour une meilleure intégration dans le web des données, l'insertion directement dans les données de base de liens vers des référentiels externes est une possibilité à considérer avec attention. Cela peut se faire soit lors de la saisie de nouvelles notices, soit de manière automatique sur l'ensemble des ressources. Les référentiels devraient pour cela être sélectionnés soigneusement par la coordination du réseau.

Dans cette même optique, l'évolution des fichiers d'autorités serait un atout. Plutôt que de les considérer et de les gérer comme de simples points d'accès aux notices

bibliographiques, on pourrait en étendre le concept afin de créer des données de référence dont les bibliothèques seraient responsables. Ceci impliquerait l'utilisation de plus de champs contrôlés, par exemple afin de saisir le pays ou la langue d'une personne, comme le fait actuellement la DNB. Suivant le modèle allemand, la fusion des notices d'autorité auteur et sujet est à entreprendre. Elle apporterait non seulement une cohérence accrue des données de RERO à l'interne comme à l'externe – notamment en LOD –, mais générerait également des gains de temps pour les catalogueurs, qui n'auraient plus qu'un fichier à gérer.

Enfin, la gestion des types de documents pourrait être repensée. Lors de la modélisation et du mapping, des incohérences particulièrement flagrantes ont été mises en évidence. Selon les règles AACR2, les données de RERO contiennent des informations concernant le type de document au sein de quatre zones différentes, dont la zone de titre, qui possède un sous-champ réservé à l'*indication générale du type de document*. Cette dernière « manque à ce point de rigueur et de cohérence qu'elle n'est pour ainsi dire d'aucune utilité en tant que critère de tri dans un catalogue », selon Philippe Le Pape (2014b). En général, ces quatre zones ne permettent pas d'indiquer plusieurs types de documents, par exemple *image* et *texte* pour un album pour enfants. Cette problématique a gagné en importance avec l'arrivée du web et de nouveaux supports d'informations, souvent hybrides. Les nouvelles règles de catalogage RDA gèrent ce problème en distinguant, au moyen de sous-champs répétitifs, trois facettes des types de documents : type de contenu, de média et de support matériel (détails au chapitre 2.2.2.2). L'adoption de RDA permettrait de résoudre des difficultés de ce genre.

6.3 Difficultés rencontrées

Bon nombre de difficultés rencontrées sont inhérentes aux données de base, gérées au moyen d'un format et de règles de catalogage vieillissants. Le chapitre précédent en donne quelques exemples. De bonnes connaissances du format MARC21 et des règles AACR2 peuvent alors se révéler très utiles.

Les données de base ont notamment rendu l'exercice de modélisation plus compliqué que prévu. L'identification des principales entités parmi les divers types de notices d'autorité ne s'est pas avérée évidente. Ce travail a soulevé la problématique inattendue de la distinction entre choses du monde réel et choses abstraites (concepts), nécessitant une analyse approfondie. Les choix effectués pour la création du modèle RERO ont donc souvent dû obéir aux contraintes des données de base, à

l'instar de leur modèle plat ou de l'éclatement des notices d'autorité. Pour échapper à ces contraintes, d'importants travaux devraient être entrepris (FRBRisation, fusion de notices), ce qui n'a pas pu se faire dans ce projet.

Les diverses étapes de modélisation ont par ailleurs exigé de nombreuses recherches et l'acquisition de nouvelles connaissances. Cette mise à jour des compétences a spécialement touché des domaines techniques tels que les données de provenance ou la structure des IRIs.

L'environnement mouvant du web sémantique en pleine émergence a également représenté une difficulté dans le cadre de ce projet. Il est en effet délicat d'effectuer des choix lorsqu'aucun standard n'est établi et que les technologies sont encore en développement. Ceci concerne par exemple les données de provenance, le choix des ontologies ou la structure du modèle. L'analyse comparative des données RDF d'autres bibliothèques s'avère alors être une solution de substitut pour s'orienter vers les standards susceptibles d'être adoptés dans le futur par la communauté.

7. Conclusion

Le web sémantique et l'Open Data représentent deux nouvelles tendances dont l'impact sur les métadonnées peut être énorme. Dans le monde entier, la communauté des bibliothèques se penche intensivement sur cette thématique, en développant de nouveaux standards pour remplacer ceux qui ont été créés il y a plus de 40 ans.

Afin de prendre part à ce mouvement, le Réseau des bibliothèques de Suisse occidentale (RERO) a lancé ce projet de publication de ses métadonnées en LOD. Le processus de modélisation et de mapping a abouti à la création d'un modèle de données qui se décline en six tables de correspondances et environ 130 règles de conversion, réutilisant exclusivement des ontologies existantes.

Selon ce modèle, les métadonnées de RERO respecteront les quatre principes des Linked Data ainsi que les cinq étoiles des LOD (cf. chapitres 2.1.2 et 2.1.3) proposés par Tim Berners-Lee (2010) :

- Elles seront désignées par des IRIs et seront déréférençables.
- Elles seront mises à disposition sous licence ouverte (Creative Commons Zero).
- Elles seront structurées pour le traitement automatique par des machines et décrites dans un format non-propriétaire (Turtle, RDF/XML, etc.) selon le standard RDF du W3C.
- Elles seront reliées à des IRIs externes de référentiels du web sémantique.

La prochaine étape logique du projet est donc la publication. Il s'agira ensuite de mettre à jour régulièrement les données en ligne, ainsi que les technologies et les ontologies utilisées. Ce processus devra être entièrement intégré aux tâches courantes du réseau.

Sur la base des LOD pourront alors être développées de nouvelles fonctionnalités pour les utilisateurs.

> *« We will shift from keyword-based searches to entity-based navigation and discovery* [...]*. So when users search for an author, for example, they will receive a summary of all the data available on the wider web, connected to our catalogue records.* [...] *We see usability as a key issue that all LOD solutions need to tackle. »* (Daniel Vila-Suero, in : The European Library 2014b)

Il est probable que les interfaces de recherche d'information soient révolutionnées. Les utilisateurs navigueront de manière différente dans les catalogues de bibliothèque. Ce changement nécessitera alors d'importants efforts dans le développement de l'utilisabilité des nouveaux services, afin de ne pas perdre des usagers, mais au

contraire d'en gagner. Au-delà de la découverte des ressources d'une institution, les catalogues deviendront des portes d'entrée vers le web, grâce aux nouvelles interconnexions créées. Le web sera dès lors également une source d'entrées vers les catalogues. Ainsi, les ressources riches et abondantes des bibliothèques sortiront de leurs portails en silos, et intégreront pleinement le web des données.

Bibliographie

About GeoNames, 2014. *GeoNames* [en ligne]. [Consulté le 7 août 2014]. Disponible à l'adresse : http://www.geonames.org/about.html

AFNOR (éd.), 1987. Disposition des données sur bande magnétique pour l'échange d'informations bibliographiques. In : *Documentation*. 6e éd. Paris : AFNOR. pp. 571-577. NF ISO 2709 (reprod. de la norme ISO 2709 publ. en 1981). ISBN 2122344601.

ARCHER, Phil, 2013. Study on persistent URIs : with identification of best practices and recommendations on the topic for the member states and the European Commission. *Phil Archer* [en ligne]. 24 juin 2013. [Consulté le 28 mars 2014]. Disponible à l'adresse : http://philarcher.org/diary/2013/uripersistence/

BEHRENS, Renate et SCHAFFNER, Verena, 2014. The adoption of RDA in the German-speaking countries. *Faster, Smarter, and Richer (FSR): Reshaping the Library Catalogue* [en ligne]. Rome. 27 février 2014. [Consulté le 12 juillet 2014]. Disponible à l'adresse : http://eprints.rclis.org/22689/1/Presentation_Behrens_Schaffner.ppt

BERMÈS, Emmanuelle, 2013. *Le web sémantique en bibliothèque* [en ligne]. Paris : Ed. du Cercle de la librairie. Collection bibliothèques. ISBN 978-2-7654-1417-9. Disponible à l'adresse : http://www.electrelaboutique.com/ProduitECL.aspx?ean=9782765414179

BERNERS-LEE, Tim, FIELDING, Roy et MASINTER, Larry, 2005. Uniform Resource Identifier (URI): Generic Syntax (RFC 3986). *The Internet Engineering Task Force* [en ligne]. janvier 2005. [Consulté le 21 juin 2014]. Disponible à l'adresse : http://www.ietf.org/rfc/rfc3986

BERNERS-LEE, Tim, HENDLER, James et LASSILA, Ora, 2001. The semantic web. *Scientific American*. mai 2001. pp. 29-37.

BERNERS-LEE, Tim, 2010. Linked Data. *World Wide Web Consortium* [en ligne]. 2010. [Consulté le 14 juin 2014]. Disponible à l'adresse : http://www.w3.org/DesignIssues/LinkedData.html

BIZER, Christian, HEATH, Tom et BERNERS-LEE, Tim, 2009. Linked Data : the story so far. *International journal on semantic web and information systems*. 2009. Vol. 5, pp. 1-22. DOI 10.4018/jswis.2009081901. This is a preprint of a paper to appear in: Heath, T., Hepp, M., and Bizer, C. (eds.). Special Issue on Linked Data, International Journal on Semantic Web and Information Systems (IJSWIS). http://linkeddata.org/docs/ijswis-special-issue

BNF, 2013a. RDA (Ressources: Description et Accès). *Bibliothèque nationale de France* [en ligne]. 16 décembre 2013. [Consulté le 12 juillet 2014]. Disponible à l'adresse : http://www.bnf.fr/fr/professionnels/rda/s.rda_objectifs.html

BNF, 2013b. RDA en France. *Bibliothèque nationale de France* [en ligne]. 16 décembre 2013. [Consulté le 28 août 2014]. Disponible à l'adresse :

http://www.bnf.fr/fr/professionnels/rda/s.rda_en_france.html?first_Art=non

BNF, 2014a. L'invité du mois : Emmanuelle Bermès. *Actualités du catalogue : lettre d'information de la BnF* [en ligne]. avril 2014. N° 34. [Consulté le 13 juin 2014]. Disponible à l'adresse : http://multimedia.bnf.fr/lettres/produits/produits34.html

BNF, 2014b. *data.bnf.fr* [en ligne]. 2014. [Consulté le 16 août 2014]. Disponible à l'adresse : http://data.bnf.fr/

BNF, 2014c. RAMEAU: Répertoire d'autorité-matière encyclopédique et alphabétique unifié. *Bibliothèque nationale de France* [en ligne]. 2014. [Consulté le 7 août 2014]. Disponible à l'adresse : http://rameau.bnf.fr/

BROWNE, Glenda et JERMEY, Jonathan, 2001. *Website indexing : enhancing access to information within websites*. Adelaide : Auslib Press. ISBN 1875145486.

CENTRE POMPIDOU, 2014. *Centre Pompidou Virtuel* [en ligne]. 2014. [Consulté le 16 août 2014]. Disponible à l'adresse : http://www.centrepompidou.fr/

CHIGNARD, Simon, 2012. *Comprendre l'ouverture des données publiques*. FYP Editions. Collection entreprendre. ISBN 9782916571706.

COYLE, Karen, 2006. Murdering MARC. *Coyle's InFormation* [en ligne]. 1 septembre 2006. [Consulté le 30 juin 2014]. Disponible à l'adresse : http://kcoyle.blogspot.ch/2006/09/murdering-marc.html

COYLE, Karen, 2013. FRBR and schema.org. *Coyle's InFormation* [en ligne]. 29 juin 2013. [Consulté le 9 juillet 2014]. Disponible à l'adresse : http://kcoyle.blogspot.ch/2013/06/frbr-and-schemaorg.html

CYGANIAK, Richard, 2014. *Prefix.cc* [en ligne]. 2014. [Consulté le 14 août 2014]. Disponible à l'adresse : http://prefix.cc/

DENTON, William, 2006. More relationships between groups. *The FRBR blog* [en ligne]. 25 février 2006. [Consulté le 11 juillet 2014]. Disponible à l'adresse : http://www.frbr.org/2006/02/25/more-relationships-between-groups

DUNSIRE, Gordon, HARPER, Corey, HILLMANN, Diane et PHIPPS, Jon, 2012. Linked Data vocabulary management: infrastructure support, data Integration, and interoperability. *Information standards quarterly*. 2012. Vol. 24, n° 2/3, pp. 4-13. DOI http://dx.doi.org/10.3789/isqv24n2-3.2012.02.

DÜRST, Martin J. et SUIGNARD, Michel, 2005. Internationalized Resource Identifiers (IRIs) (RFC 3987). *The Internet Engineering Task Force* [en ligne]. janvier 2005. [Consulté le 10 juillet 2014]. Disponible à l'adresse : http://www.ietf.org/rfc/rfc3987

ECMA INTERNATIONAL, 2013. *The JSON data interchange format* [en ligne]. octobre 2013. ECMA International. [Consulté le 10 juillet 2014]. Disponible à l'adresse : http://www.ecma-international.org/publications/files/ECMA-ST/ECMA-404.pdf

FÜRSTE, Fabian M., 2011. *Linked Open Library Data: bibliographische Daten und ihre Zugänglichkeit im Web der Daten*. Wiesbaden : Dinges & Frick. BIT online innovativ,

Bd. 33. ISBN 9783934997363.

GANDON, Fabien, 2013. Quand le lien fait sens. *Blend web mix* [en ligne]. Lyon. 2 octobre 2013. [Consulté le 4 août 2014]. Disponible à l'adresse : http://fr.slideshare.net/fabien_gandon/quand-le-lien-fait-sens

GEIPEL, Markus Michael, 2012. *Metamorph user guide* [en ligne]. 23 novembre 2012. Culturegraph. [Consulté le 29 juillet 2014]. Disponible à l'adresse : http://sourceforge.net/p/culturegraph/code/1691/tree/metamorph/trunk/docs/user_guide/metamorph.pdf?format=raw

GREAT BRITAIN. CHIEF TECHNOLOGY OFFICER COUNCIL, 2009. Designing URI sets for the UK public sector. *Gov.uk* [en ligne]. 2009. [Consulté le 10 mai 2014]. Disponible à l'adresse : https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/60975/designing-URI-sets-uk-public-sector.pdf

HARTIG, Olaf et ZHAO, Jun, 2010. Publishing and consuming provenance metadata on the web of Linked Data. In : *Provenance and annotation of data and processes : third International Provenance and Annotation Workshop, IPAW 2010, Troy, NY, USA, June 15-16, 2010, revised selected papers* [en ligne]. Berlin : Springer. juin 2010. pp. 78-90. [Consulté le 21 mai 2014]. Lecture notes in computer science. ISBN 9783642178191. Disponible à l'adresse : http://olafhartig.de/files/HartigZhao_Provenance_IPAW2010_Preprint.pdf

HÓRA, Bill de, 2007. Vocabulary design and integration. *Bill de hÓra* [en ligne]. 8 avril 2007. [Consulté le 5 août 2014]. Disponible à l'adresse : http://www.dehora.net/journal/2007/04/data_integration.html

HÜGI, Jasmin et PRONGUÉ, Nicolas, 2013. Marc contre Elodie, ou les avantages des Linked Open Data en bibliothèque. *Recherche d'ID* [en ligne]. 10 décembre 2013. [Consulté le 8 août 2014]. Disponible à l'adresse : http://recherchemid.wordpress.com/2013/12/10/marc-contre-elodie/

IFLA, 2010. *Fonctionnalités requises des données d'autorité : un modèle conceptuel* [en ligne]. Paris : BnF. [Consulté le 11 juillet 2014]. Edition française établie par la Bibliothèque nationale de France. Disponible à l'adresse : http://www.ifla.org/files/assets/cataloguing/frad/frad_2009-fr.pdf

IFLA, 2012a. *Fonctionnalités requises des notices bibliographiques: rapport final* [en ligne]. Paris : BnF. [Consulté le 11 juillet 2014]. 2e édition française établie par la Bibliothèque nationale de France. Disponible à l'adresse : http://www.ifla.org/files/assets/cataloguing/frbr/frbr-fr_2012.pdf

IFLA, 2012b. *Fonctionnalités requises des données d'autorité matière (FRSAD): un modèle conceptuel* [en ligne]. Paris : BnF. [Consulté le 11 juillet 2014]. Edition française établie par la Bibliothèque nationale de France. Disponible à l'adresse : http://www.bnf.fr/documents/frsad_rapport_final.pdf

ISELE, Robert, JENTZSCH, Anja, BIZER, Christian, VOLZ, Julius et PETROVSKI, Petar, 2014. Silk: a link discovery framework for the web of data. *Universität Mannheim* [en ligne]. 21 février 2014. [Consulté le 16 août 2014]. Disponible à l'adresse :

http://wifo5-03.informatik.uni-mannheim.de/bizer/silk/

JENTZSCH, Anja, CYGANIAK, Richard et BIZER, Christian, 2011. State of the LOD cloud. *LOD Cloud* [en ligne]. 19 septembre 2011. [Consulté le 24 juin 2014]. Disponible à l'adresse : http://lod-cloud.net/state/

KIORGAARD, Deirdre, 2009. Resource Description and Access. *National Library of Australia staff papers* [en ligne]. 2009. [Consulté le 12 juillet 2014]. Disponible à l'adresse : https://www.nla.gov.au/openpublish/index.php/nlasp/article/viewFile/1420/1724

KLEE, Carsten, 2013. Vokabulare für bibliographische Daten: zwischen Dublin Core und bibliothekarischem Anspruch. In : *(Open) Linked Data in Bibliotheken* [en ligne]. Berlin : De Gruyter Saur. pp. 45-63. Bibliotheks- und Informationspraxis, 50. ISBN 9783110276343. Disponible à l'adresse : http://dx.doi.org/10.1515/9783110278736

KULTTUURISAMPO, 2014. *Kulttuurisampo* [en ligne]. 2014. [Consulté le 16 août 2014]. Disponible à l'adresse : http://www.kulttuurisampo.fi/

LE PAPE, Philippe, 2013. FRBR : de l'expression bordel! *RDA@ABES* [en ligne]. 30 août 2013. [Consulté le 31 juillet 2014]. Disponible à l'adresse : http://rda.abes.fr/2013/08/30/frbr-de-lexpression-bordel/

LE PAPE, Philippe, 2014a. RDA, le code de catalogage qui fait grossir. *RDA@ABES* [en ligne]. 16 juin 2014. [Consulté le 31 juillet 2014]. Disponible à l'adresse : http://rda.abes.fr/2014/06/16/rda-le-code-de-catalogage-qui-fait-grossir/

LE PAPE, Philippe, 2014b. La zone zéro. *RDA@ABES* [en ligne]. 26 mai 2014. [Consulté le 9 août 2014]. Disponible à l'adresse : http://rda.abes.fr/2014/05/26/la-zone-zero/

LERESCHE, Françoise, 2004. Les formats MARC. *Ecole thématique « Documentation en mathématiques »* [en ligne]. Luminy. 15 octobre 2004. [Consulté le 24 juin 2014]. Disponible à l'adresse : http://www.rnbm.org/rencontres_2004/leresche-marc.pdf

LIBRARY LINKED DATA INCUBATOR GROUP (W3C), 2011. Datasets, values, vocabularies, and metadata element sets. *World Wide Web Consortium* [en ligne]. 25 octobre 2011. [Consulté le 10 juillet 2014]. Disponible à l'adresse : http://www.w3.org/2005/Incubator/lld/XGR-lld-vocabdataset-20111025/

LIBRARY OF CONGRESS, 2012. *Bibliographic Framework as a web of data: Linked Data model and supporting services* [en ligne]. 21 novembre 2012. Library of Congress. [Consulté le 18 juillet 2014]. Disponible à l'adresse : http://www.loc.gov/bibframe/pdf/marcld-report-11-21-2012.pdf

LIBRARY OF CONGRESS, 2014a. MARC standards. *Library of Congress* [en ligne]. 22 mai 2014. [Consulté le 24 juin 2014]. Disponible à l'adresse : http://www.loc.gov/marc/

LIBRARY OF CONGRESS, 2014b. BIBFRAME - Bibliographic Framework Initiative. *Library of Congress* [en ligne]. 2014. [Consulté le 18 juillet 2014]. Disponible à

l'adresse : http://www.loc.gov/bibframe/

MALMSTEN, Martin, 2009. Exposing library data as Linked Data. *Satellite meetings IFLA 2009: Emerging trends in technology, : libraries between web 2.0, semantic web and search technology* [en ligne]. Florence. 19 août 2009. [Consulté le 25 juillet 2014]. Disponible à l'adresse : http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.181.860&rep=rep1&type=pdf

MALMSTEN, Martin, 2013. Decentralisation, distribution, disintegration: towards Linked Data as a first class citizen in libraryland. [en ligne]. SWIB 2013. Hambourg. 27 novembre 2013. [Consulté le 8 août 2014]. Disponible à l'adresse : http://fr.slideshare.net/geckomarma/swib13/17 (diaporama) et http://www.scivee.tv/node/61571 (son)

MCCALLUM, Sally, 2010. Machine readable cataloging (MARC): 1975-2007. In : *Encyclopedia of library and information sciences*. 3rd. Boca Raton : CRC Press. pp. 3530-3539. ISBN 9780849396731.

MCGUINNESS, Deborah L., 2003. Ontologies come of age. In : *Spinning the semantic web*. Cambridge, Mass. : The MIT Press. pp. 171-194. ISBN 0262062321.

MELO, Gerard de, 2014. *Lexvo.org* [en ligne]. 2014. [Consulté le 9 août 2014]. Disponible à l'adresse : http://www.lexvo.org/

MILLER, Eric, 2013. BIBFRAME community profiles. *LC Bibliographic Framework Initiative Update Forum* [en ligne]. Chicago. 30 juin 2013. [Consulté le 18 juillet 2014]. Disponible à l'adresse : http://de.slideshare.net/zepheiraorg/alaBIBFRAME-lc20130630

MOREIRA, Miguel, 2013. RERO: les jalons posés vers le Linked Open Data. *Journée d'étude « Les données en bibliothèque - les enjeux des Linked Open Data »* [en ligne]. Lausanne. 1 octobre 2013. [Consulté le 9 décembre 2013]. Disponible à l'adresse : http://campus.hesge.ch/id_bilingue/doc/rero_lod_heg_20131001.pdf

OCLC, 2014a. OCLC releases WorldCat works as Linked Data. *OCLC* [en ligne]. 28 avril 2014. [Consulté le 15 août 2014]. Disponible à l'adresse : https://oclc.org/news/releases/2014/201414dublin.en.html

OCLC, 2014b. VIAF. *OCLC* [en ligne]. 4 avril 2014. [Consulté le 7 août 2014]. Disponible à l'adresse : http://oclc.org/viaf.en.htmlThe Virtual International Authority File

OPEN KNOWLEDGE FOUNDATION, 2012. *Le manuel de l'opendata* [en ligne]. Release 1.0.0. Open Knowledge Foundation. [Consulté le 13 juin 2014]. Disponible à l'adresse : http://opendatahandbook.org/fr/

POHL, Adrian et DANOWSKI, Patrick, 2013. Linked Open Data in der Bibliothekswelt: Grundlagen und Überblick. In : *(Open) Linked Data in Bibliotheken* [en ligne]. Berlin : De Gruyter Saur. pp. 1-44. Bibliotheks- und Informationspraxis, 50. [Consulté le 14 juin 2014]. ISBN 9783110276343. Disponible à l'adresse : http://dx.doi.org/10.1515/9783110278736

POHL, Adrian, 2014. Name authority files & Linked Data. *Übertext* [en ligne]. 10 juillet

2014. [Consulté le 6 août 2014]. Disponible à l'adresse : http://www.uebertext.org/2014/07/name-authority-files-linked-data.html

RERO, 2012. Plan stratégique RERO 2013-2017. *RERO* [en ligne]. 29 novembre 2012. [Consulté le 9 décembre 2013]. Disponible à l'adresse : http://www.rero.ch/pdfview.php?section=infos&filename=plan_strategique_rero_2013_2017.pdf

RERO, 2014a. Rapport d'activités, perspectives 2013-2014. *RERO* [en ligne]. 2014. [Consulté le 27 août 2014]. Disponible à l'adresse : https://www.rero.ch/pdfview.php?section=communique&filename=rero_rapport_activites_2013.pdf

RERO, 2014b. *RERO* [en ligne]. 2014. [Consulté le 12 août 2014]. Disponible à l'adresse : http://www.rero.ch/

RERO, 2014c. Les réservoirs de métadonnées RERO. *RERO* [en ligne]. 7 juillet 2014. [Consulté le 27 août 2014]. Disponible à l'adresse : http://data.rero.ch/

SVENSSON, Lars G., 2013. Are current bibliographic models suitable for integration with the web? *Information standards quarterly*. 2013. Vol. 25, n° 4, pp. 7-13. DOI 10.3789/isqv25no4.2013.02.

TENNANT, Roy, 2002. MARC must die. *Library Journal* [en ligne]. 15 octobre 2002. [Consulté le 24 juin 2014]. Disponible à l'adresse : http://lj.libraryjournal.com/2002/10/ljarchives/marc-must-die/

THE EUROPEAN LIBRARY, 2014a. Case study: the National Library of the Netherlands pursues its LOD strategic vision. *The European Library* [en ligne]. 9 mai 2014. [Consulté le 16 août 2014]. Disponible à l'adresse : http://www.theeuropeanlibrary.org/tel4/newsitem/5550

THE EUROPEAN LIBRARY, 2014b. Case study: National Library of Spain transforms the user experience of Linked Open Data. *The European Library* [en ligne]. 16 mai 2014. [Consulté le 12 août 2014]. Disponible à l'adresse : http://www.theeuropeanlibrary.org/tel4/newsitem/5800

UDC CONSORTIUM, 2014. UDC summary Linked Data. *UDC data* [en ligne]. 2014. [Consulté le 7 août 2014]. Disponible à l'adresse : http://udcdata.info/

VATANT, Bernard et VANDENBUSSCHE, Pierre-Yves, 2014. *Linked Open Vocabularies* [en ligne]. 12 août 2014. [Consulté le 14 août 2014]. Disponible à l'adresse : http://lov.okfn.org/

VILA-SUERO, Daniel et GÓMEZ-PÉREZ, Asunción, 2013. datos.bne.es and MARiMbA: an insight into Library Linked Data. *Library hi tech*. 2013. Vol. 31, n° 4, pp. 575-601. DOI 10.1108/LHT-03-2013-0031.

VILLAZÓN-TERRAZAS, Boris, VILCHES-BLÁZQUEZ, Luis M., CORCHO, Oscar et GÓMEZ-PÉREZ, Asunción, 2011. Methodological guidelines for publishing government Linked Data. In : *Linking government data* [en ligne]. New York, NY : Springer. pp. 27-49. [Consulté le 7 août 2014]. ISBN 978-1-4614-1766-8. Disponible à l'adresse : https://www.lri.fr/~hamdi/datalift/tuto_inspire_2012/Suggestedreadings/egovld.pdf

W3C, 2008. Cool URIs for the semantic web. *World Wide Web Consortium* [en ligne]. 3 décembre 2008. [Consulté le 16 mai 2014]. Disponible à l'adresse : http://www.w3.org/TR/cooluris/

W3C, 2009. W3C semantic web frequently asked questions. *World Wide Web Consortium* [en ligne]. 12 novembre 2009. [Consulté le 13 juin 2014]. Disponible à l'adresse : http://www.w3.org/2001/sw/SW-FAQ#swactivity

W3C, 2010. Provenance XG final report. *World Wide Web Consortium* [en ligne]. 8 décembre 2010. [Consulté le 1 août 2014]. Disponible à l'adresse : http://www.w3.org/2005/Incubator/prov/XGR-prov-20101214/

W3C, 2013a. W3C data activity : building the web of data. *World Wide Web Consortium* [en ligne]. 2013. [Consulté le 13 juin 2014]. Disponible à l'adresse : http://www.w3.org/standards/semanticweb/

W3C, 2013b. SPARQL 1.1 Overview. *World Wide Web Consortium* [en ligne]. 21 mars 2013. [Consulté le 23 juin 2014]. Disponible à l'adresse : http://www.w3.org/TR/2013/REC-sparql11-overview-20130321/

W3C, 2013c. Linked Data glossary. *World Wide Web Consortium* [en ligne]. 27 juin 2013. [Consulté le 19 juin 2014]. Disponible à l'adresse : http://www.w3.org/TR/2013/NOTE-ld-glossary-20130627/

W3C, 2013d. Ontologies. *World Wide Web Consortium* [en ligne]. 2013. [Consulté le 24 juin 2014]. Disponible à l'adresse : http://www.w3.org/standards/semanticweb/ontology

W3C, 2013e. Linked Data cookbook. *World Wide Web Consortium* [en ligne]. 15 mars 2013. [Consulté le 16 août 2014]. Disponible à l'adresse : http://www.w3.org/2011/gld/wiki/Linked_Data_Cookbook

W3C, 2014a. RDF 1.1 primer. *World Wide Web Consortium* [en ligne]. 25 février 2014. [Consulté le 21 juin 2014]. Disponible à l'adresse : http://www.w3.org/TR/2014/NOTE-rdf11-primer-20140225/

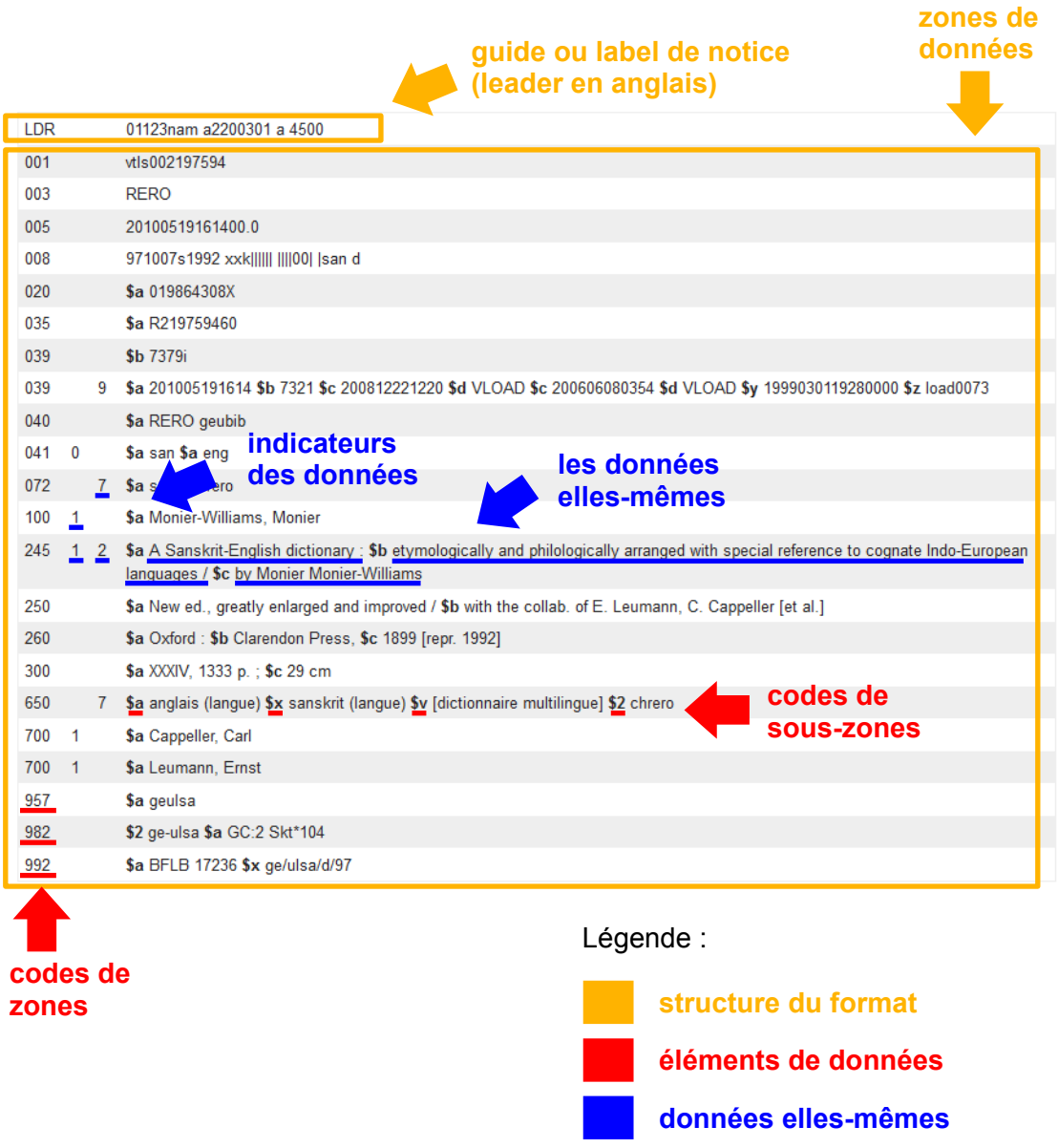
W3C, 2014b. Best practices for publishing Linked Data. *World Wide Web Consortium* [en ligne]. 9 janvier 2014. [Consulté le 7 août 2014]. Disponible à l'adresse : http://www.w3.org/TR/2014/NOTE-ld-bp-20140109/

WALLIS, Richard, 2013. Schema Bib Extend. *Information standards quarterly*. 2013. Vol. 25, n° 4, pp. 30-32. DOI 10.3789/isqv25no4.2013.06.

WILLER, Mirna (éd.), 2009. *UNIMARC manual: authorities format*. 3rd ed. München : K.G. Saur. IFLA series on bibliographic control, 38. ISBN 9783598242861.

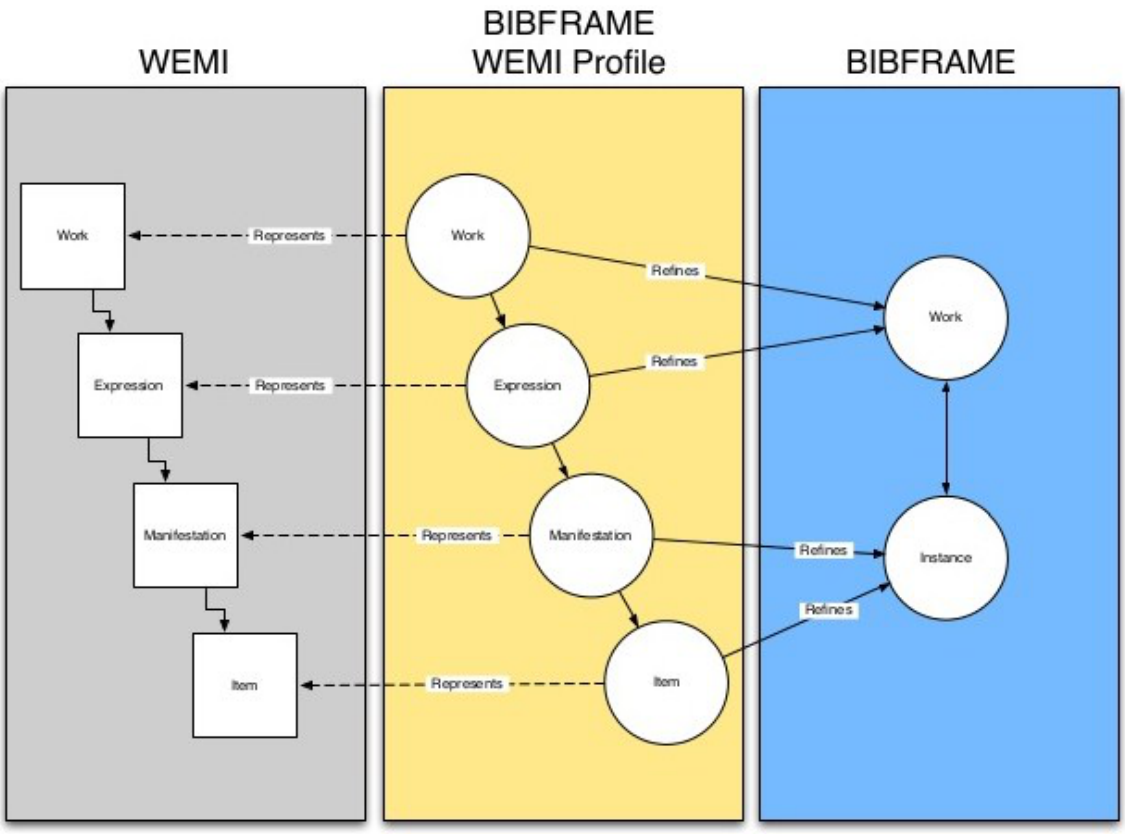
Annexe 1 : Structure d'une notice MARC

Dans la notice expliquée ci-dessous, le répertoire ainsi que le séparateur de notice ne sont pas visibles.



LDR			01123nam a2200301 a 4500
001			vtls002197594
003			RERO
005			20100519161400.0
008			971007s1992 xxk|||||| ||||00| |san d
020			**$a** 019864308X
035			**$a** R219759460
039			**$b** 7379i
039		9	**$a** 201005191614 **$b** 7321 **$c** 200812221220 **$d** VLOAD **$c** 200606080354 **$d** VLOAD **$y** 1999030119280000 **$z** load0073
040			**$a** RERO geubib
041	0		**$a** san **$a** eng
072		7	**$a** s… …ero
100	1		**$a** Monier-Williams, Monier
245	1	2	**$a** A Sanskrit-English dictionary : **$b** etymologically and philologically arranged with special reference to cognate Indo-European languages / **$c** by Monier Monier-Williams
250			**$a** New ed., greatly enlarged and improved / **$b** with the collab. of E. Leumann, C. Cappeller [et al.]
260			**$a** Oxford : **$b** Clarendon Press, **$c** 1899 [repr. 1992]
300			**$a** XXXIV, 1333 p. ; **$c** 29 cm
650		7	**$a** anglais (langue) **$x** sanskrit (langue) **$v** [dictionnaire multilingue] **$2** chrero
700	1		**$a** Cappeller, Carl
700	1		**$a** Leumann, Ernst
957			**$a** geulsa
982			**$2** ge-ulsa **$a** GC:2 Skt*104
992			**$a** BFLB 17236 **$x** ge/ulsa/d/97

Annexe 2 : Profil de communauté BIBFRAME pour FRBR



(Miller 2013, p. 21)

Annexe 3 : Exemple de description VoID

La description VoID ci-dessous, en format RDF/XML, concerne le catalogue collectif RERO. Etant provisoire, elle contient des séries de X pour les données encore inconnues lors de sa conception.

```
<?xml version="1.0"?>

<rdf:RDF
      xmlns:cc="http://creativecommons.org/ns#"
      xmlns:dct="http://purl.org/dc/terms/"
      xmlns:foaf="http://xmlns.com/foaf/0.1/"
      xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
      xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
      xmlns:void="http://rdfs.org/ns/void#">

      <rdf:Description rdf:about="http://data.rero.ch/catalogue_collectif/about">
            <rdf:type rdf:resource="http://rdfs.org/ns/void#DatasetDescription"/>
            <foaf:primaryTopic
rdf:resource="http://data.rero.ch/catalogue_collectif"/>
            <dct:creator rdf:resource="http://data.rero.ch/02-A005399379"/>
            <dct:created rdf:datatype="http://www.w3.org/2001/XMLSchema#date">2014-
XX-XXTXX:XX:XX+01:00</dct:created>
            <dct:modified
rdf:datatype="http://www.w3.org/2001/XMLSchema#dateTime">2014-XX-
XXTXX:XX:XX+01:00</dct:modified>
            <dct:issued
rdf:datatype="http://www.w3.org/2001/XMLSchema#dateTime">2014-XX-
XXTXX:XX:XX+01:00</dct:issued>
            <rdfs:label>Document that describes the dataset of the Union catalog of
RERO</rdfs:label>
      </rdf:Description>

      <rdf:Description rdf:about="http://data.rero.ch/catalogue_collectif">
            <rdf:type rdf:resource="http://rdfs.org/ns/void#Dataset"/>
            <void:dataDump rdf:resource="http://data.rero.ch/XXX"/>
            <void:exampleResource rdf:resource="http://data.rero.ch/01-R219759460"/>
            <void:sparqlEndpoint>http://data.rero.ch/XXX</void:sparqlEndpoint>
            <void:uriSpace>http://data.rero.ch/</void:uriSpace>
            <void:vocabulary rdf:resource="http://purl.org/ontology/bibo/"/>
            <void:vocabulary rdf:resource="http://purl.org/dc/elements/1.1/"/>
            <void:vocabulary rdf:resource="http://purl.org/dc/terms/"/>
            <void:vocabulary rdf:resource="http://www.europeana.eu/schemas/edm/"/>
            <void:vocabulary rdf:resource="http://xmlns.com/foaf/0.1/"/>
            <void:vocabulary rdf:resource="http://rdaregistry.info/Elements/u/"/>
            <void:vocabulary  rdf:resource="http://www.w3.org/1999/02/22-rdf-syntax-
ns# "/>
            <void:vocabulary  rdf:resource="http://www.w3.org/2000/01/rdf-schema#"/>

            <void:vocabulary rdf:resource="http://www.w3.org/1999/02/22-rdf-syntax-
ns#"/>
            <dct:title>Catalogue collectif RERO</dct:title>
            <dct:description>"Catalogue collectif RERO" is the union catalog of
public, academic and patrimonial libraries of the French-speaking part of
Switzerland.</dct:description>
            <dct:creator rdf:resource="http://data.rero.ch/02-A005399379"/>
            <dct:license
rdf:resource="http://creativecommons.org/publicdomain/zero/1.0/"/>
            <cc:license
rdf:resource="http://creativecommons.org/publicdomain/zero/1.0/"/>
            <rdfs:isDefinedBy
rdf:resource="http://data.rero.ch/catalogue_collectif/about"/>
      </rdf:Description>

</rdf:RDF>
```

Annexe 4 : Exemple de données de provenance

Ces données de provenance, en format RDF/XML, concernent la ressource <http://data.rero.ch/01-R219759460>.

```
<?xml version="1.0"?>

<rdf:RDF
      xmlns:cc="http://creativecommons.org/ns#"
      xmlns:dct="http://purl.org/dc/terms/"
      xmlns:foaf="http://xmlns.com/foaf/0.1/"
      xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
      xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
      xmlns:void="http://rdfs.org/ns/void#">

      <rdf:Description rdf:about="http://data.rero.ch/01-R219759460/about">
            <foaf:primaryTopic rdf:resource="http://data.rero.ch/01-R219759460"/>
            <dct:creator rdf:resource="http://data.rero.ch/02-A005399379"/>
            <dct:created
rdf:datatype="http://www.w3.org/2001/XMLSchema#dateTime">1998-11-
02T16:15:05+01:00</dct:created>
            <dct:modified
rdf:datatype="http://www.w3.org/2001/XMLSchema#dateTime">2010-05-
31T11:54:19+01:00</dct:modified>
            <dct:issued
rdf:datatype="http://www.w3.org/2001/XMLSchema#dateTime">2014-08-
31T23:59:00+01:00</dct:issued>
            <void:inDataset rdf:resource="http://data.rero.ch/catalogue_collectif"/>
            <rdf:type rdf:resource="http://purl.org/ontology/bibo/Document"/>
      </rdf:Description>

</rdf:RDF>
```

Explications :

- *foaf:primaryTopic* établit le lien entre la notice *about* et la ressource réelle.
- *dct:creator* indique le créateur des données (RERO en l'occurrence).
- *dct:created* indique la date de création des données de base (en format MARC21).
- *dct:modified* indique la date de dernière modification des données de base.
- *dct:issued* indique la date de publication des données RDF.
- *void:inDataset* établit le lien entre la notice *about* et le jeu de données
- *rdf:type* indique que cette notice *about* fait partie de la classe *bibo:Document*.