Impact des données ouvertes et liées sur les catalogues bibliographiques

Fabien Duchateau [1], **Nicolas Lumineau** [1], **Trond Aalberg** [2]

1. *LIRIS, UMR5205, Université Claude Bernard Lyon 1, Université de Lyon, Lyon, France*
prénom.nom@liris.cnrs.fr

2. *NTNU*
Trondheim, Norvège
prénom.nom@idi.ntnu.no

RÉSUMÉ. *Les systèmes d'information des bibliothèques incluent un élément central, le catalogue, dont les notices bibliographiques décrivent chacune des œuvres disponibles (e.g., livre, enregistrement vidéo). La gestion de ce catalogue a peu évolué au cours des dernières décennies et rencontre de nombreux problèmes : un modèle plat, qui ne permet pas de représenter simplement d'autres entités et leurs relations, de nombreuses incohérences et ambiguïtés dans les données, des pratiques locales de catalogage limitant les échanges, etc. Une transition vers le web sémantique implique une transformation du catalogue en profondeur : la définition d'un nouveau modèle, la migration de données garantissant un haut degré de qualité, l'enrichissement des données originales et l'amélioration des services associés au catalogue (e.g., recherche sémantique). Pour réussir ces différentes étapes, l'exploitation des données ouvertes et liées est nécessaire, que ce soit pour faciliter l'appariement d'ontologies, pour lever des ambiguïtés avec l'aide de référentiels ou pour lier des entités vers d'autres sources. Cet article dresse donc un panorama de l'impact des données ouvertes et liées dans le domaine bibliographique, en particulier sur la migration de son catalogue vers le web sémantique.*

ABSTRACT. *Integrated library systems manage the catalog of bibliographic records. This catalog has not evolved much during the last decades, and many problems arise due to its flat model, redundancies, inconsistencies and local practices. To remain relevant in the modern computing world, libraries need to adopt a new model, to migrate the data and to provide enhanced services. Linked open data can be useful during this transition. This paper describes an overview of the impact of linked data in the bibliographic domain.*

MOTS-CLÉS : *systèmes d'information, bibliothèques numériques, SIGB, catalogue, données liées, données ouvertes, intégration de données, enrichissement sémantique.*

KEYWORDS: *integrated library systems, linked open data, data integration, semantic enrichment.*

DOI:10.3166/ISI.23.3-4.57-93

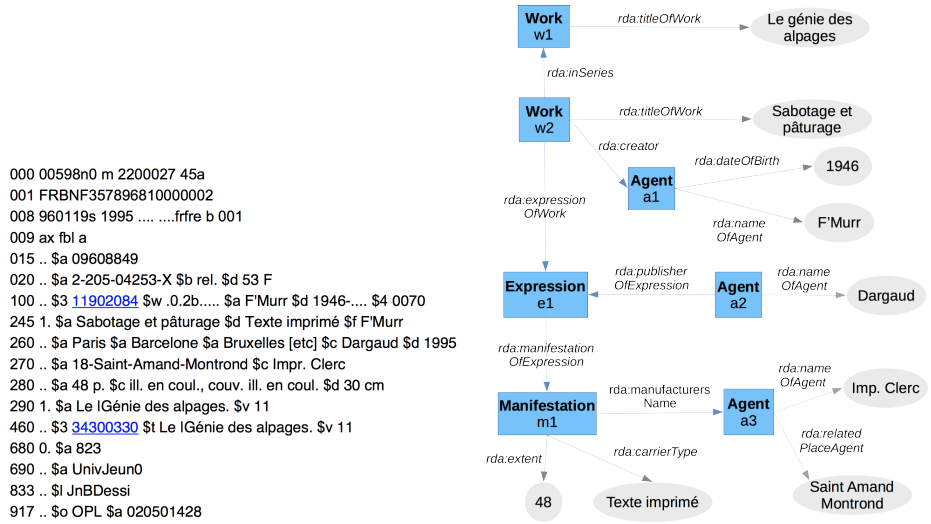
1. Introduction

Les institutions culturelles, comme les musées ou les bibliothèques, sont en charge de répertorier, de préserver, et de diffuser la culture et le patrimoine. Dans les années 1960, l'informatisation augure de nouvelles possibilités pour faciliter les missions des institutions culturelles et pour consulter et partager les éléments patrimoniaux et culturels ainsi que les informations qui les décrivent. Dans cet article, nous nous intéressons plus spécifiquement aux bibliothèques [1]. Celles-ci mettent à disposition d'un public des collections de documents (livres, enregistrements audio ou vidéo, articles scientifiques, etc.). En plus des bibliothèques publiques (municipales, universitaires), il existe de nombreuses bibliothèques privées (e.g., médicales, juridiques, musicales), et toutes doivent au minimum gérer des informations relatives aux documents. Le système intégré de gestion de bibliothèque (SIGB) est le système d'information des bibliothèques (Wang, Dawes, 2012). Il regroupe de nombreux modules, comme la gestion des usagers (abonnement), des acquisitions (commandes, factures, etc.), de circulation de documents (prêts, retours, réservations, etc.) et du catalogue. Ce dernier est un élément crucial du SIGB : il liste et décrit les documents disponibles (notices bibliographiques), autorisant ainsi la recherche, la consultation ou l'emprunt de documents. Il sert également de vitrine pour l'institution, puisque la plupart des catalogues sont intégrés dans un portail documentaire accessible sur le web (Kaenel, Iriarte, 2007).

L'informatisation des bibliothèques a nécessité le développement de catalogues informatisés, et donc d'un format de stockage et d'échange des notices bibliographiques. Le format *Machine-Readable Cataloging* (MARC) est donc défini en 1968 par la *Library of Congress* (Furrie, 2000). Il est basé sur un ensemble de champs de données, et chaque champ a pour étiquette un nombre de trois chiffres (e.g., le champ *100* pour décrire la personne créatrice de l'œuvre). Certains champs peuvent être divisés en sous-champs, dont l'étiquette se compose du symbole dollar suivi d'un autre caractère (e.g., *100$a* pour spécifier le nom de la personne créatrice, *100$d* pour ses dates de naissance et de décès). La figure 1a illustre une notice pour la bande-dessinée *le génie des alpages*, tome *sabotage et pâturage*, de *F'Murr*. Sur cette notice, nous pouvons observer 17 champs et 32 sous-champs.

Au niveau du web, un nouvel essor s'est récemment développé autour du web sémantique et des données ouvertes et liées (Berners-Lee *et al.*, 2001 ; Bizer *et al.*, 2009). Le web sémantique tend à promouvoir l'utilisation de formats de données qui facilitent le partage, la réutilisation, le traitement par des machines et qui permettent de produire de nouvelles connaissances grâce au raisonnement. Les données liées sont une méthode de publication des données qui favorisent le traitement automatisé et l'établissement de relations vers d'autres sources de données (Gandon *et al.*, 2012). Cette interconnexion de multiples sources de données facilite la combinaison d'in-

1. Bien que centré sur les bibliothèques, cette vue d'ensemble de l'impact des données liées se vérifie en grande partie pour les autres institutions culturelles telles que les musées.



(a) Au format MARC (source BNF)

(b) Selon le modèle FRBR (version simplifiée et utilisant le vocabulaire RDA)

Figure 1. Représentation d'une notice bibliographique

formations pour répondre à une requête par exemple. Si les données sont ouvertes, leur accès et leur réutilisation ne sont pas restreintes (licence ouverte). D'une manière générale, le web sémantique et les données liées n'ont pas pour objectif de lier des documents (web classique) mais plutôt de lier les objets ou entités contenus dans ces documents. Les bibliothèques ont rapidement pressenti le potentiel lié à l'exploitation du web sémantique et des données liées (Harper, Tillett, 2007 ; Malmsten, 2008 ; Goddard, Byrne, 2010).

Pour les bibliothèques, la transition vers des catalogues publiés sous forme de données ouvertes et liées (LOD, pour *Linked Open Data*) pose plusieurs problèmes, principalement au niveau de la qualité des données (Gonzales, 2014). En 2011, l'incubateur du W3C *Library Linked Data* liste dans son rapport final les problèmes que rencontre la communauté pour tendre vers cette transition [2]. Tout d'abord, le modèle de données MARC actuellement utilisé par les bibliothèques est difficilement compatible avec les données liées, que ce soit à cause de sa sémantique ambigüe ou de l'impossibilité de référencer directement une entité. Il est donc nécessaire de transformer les notices bibliographiques vers un modèle plus « sémantique », qui réutilise si possible des concepts existants dans les ontologies et vocabulaires des données liées et ouvertes. Mais la prolifération des modèles et vocabulaires ne facilite pas cette modélisation des données. Un second problème concerne la migration de données. L'absence de sémantique bien établie dans les notices impose de lever les ambiguï-

2. http://www.w3.org/2005/Incubator/lld/XGR-lld/

tés, comme par exemple préciser le rôle (dessinatrice, traductrice, etc.) des personnes secondaires mentionnées dans la notice d'une œuvre. Le passage d'un format notice (décrivant complètement une œuvre) à un format sémantique (centré sur les entités) nécessite de détecter les entités équivalentes décrites dans les notices. Un autre défi consiste à identifier et interpréter les motifs bibliographiques principaux (e.g., traduction, agrégation de plusieurs œuvres, adaptation à partir d'une autre œuvre) qui apparaissent de manière implicite (ou « cachée ») dans les notices MARC. Il est évident que les données ouvertes et liées, et notamment l'exploitation de référentiels comme VIAF [3] ou RAMEAU [4], peuvent faciliter cette transformation. Puisqu'un catalogue « sémantique » isolé apporterait peu d'intérêt par rapport à sa version d'origine, il convient de le lier à d'autres sources de données, comme des référentiels et des bases de connaissances plus générales (e.g., DBpedia [5]). Le liage d'entités est très étudié en dehors de la communauté bibliographique, mais les spécificités de cette dernière nécessitent des adaptations. Enfin, les services associés au catalogue ne sont pas valorisés. Par exemple, la recherche, qui repose généralement sur un moteur de recherche fédéré, connaît un certain désintérêt, bien souvent au profit de moteurs de recherche comme Google dont les résultats sont plus pertinents et moins redondants (Gibson *et al.*, 2009 ; Chen, 2006). L'exploitation du catalogue sémantique nécessite donc d'une part, de résoudre les problèmes d'indexation, et d'autre part, d'améliorer la visualisation et l'interprétation des requêtes d'un point de vue usager de la bibliothèque.

L'objectif de cet article est de dresser un panorama de l'impact des données liées et ouvertes sur les bibliothèques, et en particulier sur la gestion du catalogue. Pour cela, nous étudierons d'abord cet impact dans le but d'obtenir un catalogue qui réponde aux besoins du web sémantique. Cela implique de s'intéresser d'abord au choix d'un modèle de données « sémantique » pour le catalogue (section 2), puis à la transformation du catalogue MARC vers ce nouveau modèle (section 3). Ensuite, nous verrons comment lier un catalogue à d'autres sources de données, et nous présenterons les jeux de données déjà disponibles sous forme de données ouvertes et liées (section 4). Enfin, nous étudierons l'impact des données liées sur l'utilisation du catalogue sémantique (section 5). La conclusion ouvrira sur les perspectives offertes par les données ouvertes et liées pour le monde bibliographique.

2. Élaboration du modèle sémantique

Bien que très répandu dans le domaine bibliographique, le format MARC présente quelques inconvénients majeurs (Tennant, 2002). Tout d'abord, des alternatives co-existent (e.g., UNIMARC, MARC21) et limitent ainsi l'échange et la réutilisation directe de notices. De plus, de nombreux champs secondaires (e.g., notes) sont utilisés pour saisir des informations importantes, mais ils ne respectent pas de convention (saisie libre). Chaque institution possède ses propres pratiques de catalogage, et la

3. http://viaf.org/
4. http://rameau.bnf.fr/
5. http://dbpedia.org/

signification d'une donnée peut donc varier d'une institution à l'autre, voire au sein de la même institution lorsque différents catalogueurs se succèdent. Enfin, une notice MARC est une description d'un objet physique, et qui contient donc des informations à différents niveaux : par exemple, le nombre de pages concerne bien l'objet physique, mais le titre de l'œuvre ou le nom de l'auteur se rapportent au travail intellectuel (identique quelque soit l'édition). La communauté a également identifié des motifs bibliographiques récurrents (e.g., traduction, adaptation d'un roman en film), que les notices MARC ne peuvent facilement représenter (Riva, 2004). En résumé, le format MARC est source de redondance, d'incohérence, d'un manque de granularité et d'une grande complexité, en particulier pour les non-bibliothécaires (Alemu *et al*., 2012). Pour pallier les problèmes de MARC, les spécifications *Functional Requirements for Bibliographic Records* (FRBR) ont été définies dans les années 1990, puis améliorées sous l'appellation *Library Reference Model* (LRM) en 2015 (Committee, Group, 1998 ; Riva *et al*., 2016). La figure 1b présente une version FRBRisée mais incomplète de la notice MARC décrivant la bande-dessinée *Sabotage et pâturage* de *F'Murr*. Les entités bibliographiques principales sont le *Work* (travail intellectuel), l'*Expression* (réalisation artistique ou intellectuelle du *Work*, comme une traduction), et la *Manifestation* (objet physique, comme un livre ou un document numérique). Les *Agents* décrivent des personnes ou organisations en lien avec les entités bibliographiques. Dans cet exemple, nous avons deux *Works* puisque la série *le génie des alpages* a autant d'importance que le tome. Notons qu'une autre entité FRBR, l'*Item*, non présente dans l'exemple, permet de représenter et décrire un exemplaire de la Manifestation (par exemple, en précisant son état, endommagé ou protégé par une couverture). Du point de vue base de données, par son côté modélisation en entité/association, FRBR se rapproche d'un modèle conceptuel. Il se situe à un niveau d'abstraction élevé, et reste indépendant des pratiques de catalogage et d'implémentation, ce qui ne le rend pas directement exploitable dans une application (Coyle, 2014 ; 2016). Des modèles implémentables ont été proposés, ils réutilisent des concepts provenant des données ouvertes et liées. Dans cette section, nous présentons brièvement ces modèles sémantiques, puis nous décrivons les défis pour choisir un modèle pertinent afin de publier un catalogue sous forme de données liées et ouvertes.

2.1. Classification des modèles sémantiques

Différentes implémentations du modèle FRBR ont été étudiées (Baker *et al*., 2014 ; Coyle, 2016). La plupart réutilisent des vocabulaires généraux, comme SKOS, Dublin Core Terms, FOAF, International Standard Bibliographic Description (ISBD) [6] ou `schema.org` [7] et son extension spécifique BiblioGraph [8], mais elles (re)définissent aussi les concepts propres au monde bibliographique. Les modèles qui possèdent un

6. http://metadataregistry.org/schema/show/id/25.html
7. http://schema.org/
8. http://bibliograph.net/

concept de Work uniquement, comme `schema.org` (classe *CreativeWork*) ou Wikidata [9] (classe *work*), ne sont pas discutés ici.

FRBRcore [10] est une première implémentation en RDF, mais avec des variantes par rapport aux spécifications de FRBR (e.g., ajout de sous-classes pour l'œuvre, mais certaines ne sont pas reconnues par le monde bibliographique). La version la plus conforme aux principes FRBR se nomme FRBRer [11] et se base sur la philosophie entité/association. Cependant, elle ne respecte pas toutes les pratiques du web sémantique (e.g., pas de relations hiérarchiques entre classes). Pour pallier les inconvénients de FRBRer, une version orientée-objet, FRBRoo [12], a été proposée en 2008. Elle étend les concepts du *CIDOC Conceptual Reference Model* [13] et permet donc de représenter des données d'autres domaines (e.g., musées). Enfin, RDA [14] (*Resource Description & Access*) contient un vocabulaire complet en RDF basé sur les règles de cataloguage des éléments de FRBR (Westrum *et al.*, 2012 ; Phipps *et al.*, 2015). Il définit également un ensemble de correspondances vers d'autres vocabulaires, notamment les *Dublin Core Metadata Terms* qui sont également très utilisés en dehors du monde bibliographique et facilitent donc la compréhension par des non-bibliothécaires.

En parallèle des modèles qui respectent (globalement) FRBR, des alternatives sont apparues, parfois avec un objectif de simplification. BIBFRAME [15] ne suit pas les spécifications de FRBR, mais s'en inspire (Kroeger, 2013). Par exemple, seulement deux classes sont présentes pour les œuvres (*Work* et *Instance*). Comme la communauté bibliographique utilise de nombreux titres, ce modèle inclut une classe générique pour ces différents titres. Le vocabulaire de BIBFRAME est compatible avec le web sémantique, et il existe un système de hiérarchie entre les classes. Les ontologies bibliotek-o [16] et LD4L [17] sont proposées par les universités de Cornell, Harvard, Iowa et Standford. Le modèle LD4L avait pour objectif de produire une ontologie simplifiée qui facilite sa réutilisation en dehors du monde bibliographique (Krafft, 2015). Il reprend les concepts de BIBFRAME, mais des propositions d'amélioration sont suggérées suite à des expérimentations sur jeux de données réelles. La version récente, bibliotek-o, est également une extension de BIBFRAME (Kovari *et al.*, 2017). Elle ajoute notamment un certain nombre de propriétés et relations ainsi que deux classes, à savoir *Activity* et *TitleElement* (ce dernier permettant un découpage plus fin de la classe Title de BIBFRAME). L'ontologie FaBiO [18] pour *FRBR-aligned Bibliographic Ontology* est fortement basée sur FRBRcore, et l'étend en ajoutant des relations entre

9. http://www.wikidata.org/
10. http://vocab.org/frbr/core
11. http://metadataregistry.org/schema/show/id/5.html
12. http://metadataregistry.org/schema/show/id/94.html
13. http://www.cidoc-crm.org/frbroo/
14. http://www.rdaregistry.info
15. https://www.loc.gov/bibframe/
16. https://bibliotek-o.org/
17. https://ld4l.org/
18. http://www.sparontologies.net/ontologies/fabio

classes (e.g., entre *Work* et *Manifestation*), ce qui facilite la modélisation de certains domaines et notamment celui des publications scientifiques. De même, l'ontologie bibliographique BIBO [19] est spécialisée dans la gestion des publications scientifiques.

Le tableau 1 synthétise les modèles généraux basés sur les principes FRBR. Tous ces modèles disposent au minimum d'une implémentation en RDF. Les premières lignes du tableau présentent des caractéristiques générales sur les modèles : année de parution de la première version, personnes ou institution responsable de la création du modèle, et type de modèle. Concernant les vocabulaires, tous les modèles réutilisent les vocabulaires « de base », à savoir OWL, RDF, RDFS, DC, DC Terms et SKOS. Ceux-ci ne sont donc pas affichés dans le tableau. De plus, chaque modèle définit également son propre vocabulaire, indiqué en italique dans le tableau. Le vocabulaire RDA *Regional Encoding* (RE) [20] permet de prendre en compte les différents types d'encodage des médias. Il est supporté par plusieurs modèles. Le modèle bibliotek-o repose sur de nombreux vocabulaires comme les standards W3C sur les annotations (OA) [21] et la provenance (PROV) [22], schema.org [23] pour les événements, VIVO [24] pour des spécificités académiques (e.g., ordre des auteurs d'une publication scientifique), et les éléments RDA sans contraintes pour des sous-classes. Les deux lignes suivantes du tableau donnent un aperçu des points forts et points faibles. FRBRcore, bien qu'extensible, est par exemple considéré comme incomplet d'un point de vue libraire, ce qui nécessite un effort d'adaptation conséquent avant de l'exploiter dans un contexte bibliographique. RDA est actuellement adopté par de nombreuses institutions (Amérique du Nord, Europe), et le caractère intuitif de FRBR peut le rendre plus accessible à des non-bibliothécaires, d'autant plus qu'il est proche des données réelles (règles de catalogage). Karen Coyle pointe comme principal point faible de RDA l'absence de scénarios et cas d'utilisation, qui limitent l'orientation à suivre pour son développement (Coyle, 2016). FRBRoo intègre les données d'institution culturelles comme les musées et apporte une gestion de déclenchements d'événements, mais son niveau élevé de détails et sa complexité peuvent limiter son adoption. FRBRer, par sa conformité aux principes FRBR, apparait comme une solution peu compatible avec le web sémantique. BIBFRAME s'éloigne le plus des principes de FRBR, en ne conservant que les classes *Work* et *Instance*. Cette vision alternative peut rendre difficile l'interopérabilité avec les autres modèles, même si des correspondances (parfois partielles) sont proposés avec MARC21 [25], avec le modèle d'Europeana EDM (Zapounidou *et al*., 2014) ou avec FRBRer (Zapounidou *et al*., 2017). Enfin, LD4L et son successeur bibliotek-o sont plutôt des modèles expérimentaux utilisés pour valider des propositions d'amélioration (e.g., certaines propositions de LD4L ont été intégrées dans la

19. http://bibliontology.com/
20. http://metadataregistry.org/vocabulary/show/id/451.html
21. https://www.w3.org/ns/oa
22. https://www.w3.org/ns/prov
23. http://schema.org/
24. http://vivoweb.org/
25. https://www.loc.gov/bibframe/mtbf/

version 2.0 de BIBFRAME). Enfin, la dernière ligne du tableau fournit des exemples d'utilisation de chaque modèle dans les bibliothèques. FRBRcore est principalement implémenté dans des institutions allemandes, notamment l'union des catalogues bibliothécaires de Bavière, Berlin et Brandenburg (B3Kat) [26] et le regroupement des bibliothèques de Rhénanie-du-Nord-Westphalie (HBZ) [27]. Par exemple, B3Kat contient 26 millions d'œuvres, pour un total de 980 millions de triplets RDF. RDA est expérimenté avec des bibliothèques universitaires, en France avec le système universitaire de documentation (SUDOC) [28], au Royaume-Uni avec RLUK [29] ou à l'université de Chicago (UCL) (Cronin, 2011). Mais de nombreuses autres institutions s'intéressent à RDA, et des groupes de travail s'organisent pour adapter RDA aux spécificités nationales (e.g., en France, Transition Bibliographique [30]). Ainsi, la bibliothèque Cervantes [31] a transformé 200 000 notices selon le modèle RDA (Candela *et al.*, 2016). FRBRoo a été implémenté par un réseau de bibliothèques publiques et académiques, DLF et NUKAT, en Pologne (Mazurek *et al.*, 2012). Il est aussi utilisé comme modèle de base, étendu avec de nouveaux concepts, dans le cadre des projets musicaux MUSES, pour ManUScript Italian poEtry in muSic (Luzzi, 2014), et DOREMUS (Choffé, Leresche, 2016). FRBRer a également fait l'objet de différentes expérimentations malgré ses limitations [32]. La bibliothèque nationale espagnole a ainsi sémantisé son catalogue selon FRBRer (Vila-Suero *et al.*, 2013). L'union des catalogues suédois (LIBRIS) [33] se base sur FRBRer mais a défini son propre vocabulaire. BIBFRAME est déjà utilisé par de nombreuses institutions [34], comme la Library of the Hungarian National Museum (HCN) [35], la bibliothèque nationale allemande (DNB) [36], et de nombreuses bibliothèques universitaires (e.g., Illinois [37], UCL [38], Naples [39]). LD4L (et dans une moindre mesure la récente proposition bibliotek-o) a démontré sa capacité à résoudre des problèmes de modélisation rencontrés notamment avec BIBFRAME. Différents projets exploitant LD4L [40] étudient la modélisation de ressources spécifiques, que ce soit pour la cartographie (McGee *et al.*, 2017) ou le domaine des arts (Billey *et al.*, 2018).

---

26. http://lod.b3kat.de/
27. http://lobid.org/
28. https://old.datahub.io/dataset/sudocfr
29. http://www.theeuropeanlibrary.org/tel4/access/data/lod
30. http://www.transition-bibliographique.fr/
31. http://data.cervantesvirtual.com/
32. Lorsque le modèle FRBR est évoqué, il s'agit en général de FRBRer. Certains travaux ne mentionnent donc pas directement FRBRer.
33. http://libris.kb.se/
34. https://www.loc.gov/bibframe/implementation/register.html
35. http://data.hnm.hu/
36. https://de.slideshare.net/sollbruchstelle/2014-0126-bibframeheuvelmann
37. http://sif.library.illinois.edu/bibframe/
38. http://www.ucl.ac.uk/dis/research/collaborativeprojects/lobd
39. http://catalogo.share-cat.unina.it/sharecat/
40. https://wiki.duraspace.org/display/LD4P/Project+Pages

Tous ces modèles sont des variantes plus ou moins proches de FRBR. Ils possèdent chacun des avantages et inconvénients, mais ont été pensés pour la publication de catalogue sous forme de données liées. Suominen et Hyvönen ont proposé une classification des différents modèles, mais d'un point de vue bibliothécaire, par exemple selon la présence ou non du concept de Work (Suominen, Hyvönen, 2017). Les auteurs critiquent l'utilisation de multiples modèles et vocabulaires dans les projets LOD, ce qui ne facilite pas l'interopérabilité entre les catalogues et restreint l'intérêt pour ces données hors du monde bibliographique.

2.2. Choix de modélisation

Des connaissances approfondies sur chaque modèle sont indispensables pour sélectionner le plus pertinent en fonction du catalogue à publier. Une étude a montré que les participants d'une expérimentation norvégienne visant à transformer une collection sous forme de données liées ont perçu la création et la maintenance de l'ontologie comme la tâche la plus pénible, notamment à cause des nombreux concepts similaires dans différents vocabulaires (Bygstad *et al*., 2009). De plus, un jeu de données publié sous forme de données liées et ouvertes de bonne qualité doit idéalement satisfaire cinq pré-requis [41], dont l'utilisation d'URI pour décrire les objets et la création de liens vers d'autres jeux de données. Pour choisir un modèle pertinent, trois étapes sont nécessaires :

– identification des concepts pertinents (sélection dans un modèle ou création);
– représentation des URI;
– sélection des thésaurus ou autres vocabulaires contrôlés de valeurs.

La modélisation consiste tout d'abord à sélectionner les concepts (classes et propriétés). Le modèle sélectionné peut ne pas inclure certains concepts, et il est nécessaire de les créer ou de les trouver dans d'autres ontologies. L'intérêt principal de réutiliser une ontologie existante, c'est d'une part d'attirer un public plus large que celui de la communauté bibliographique et d'autre part d'éviter la redondance des concepts (e.g., `dcterms:Agent` et `foaf:Agent`). Par exemple, la *British Library* a défini une propriété (`blt:bnb`) pour stocker l'identifiant de notice d'origine, et sa propre classe *publication statement*. Elle a réutilisé des concepts de l'ontologie Geonames pour les pays de création d'une œuvre ou d'une personne (Deliot, 2014). De même, dans les travaux de chercheurs espagnols, la description des événements temporels repose sur OWL-Time (Candela *et al*., 2016). Europeana regroupe des collections d'objets du patrimoine culturel (Doerr *et al*., 2010). Son modèle, *Europeana Data Model* (EDM) [42], est plus générique que ceux dédiés au monde bibliographique, mais il réutilise de nombreux concepts existants, notamment RDFS, CIDOC CRM, SKOS, FOAF, Dublin Core et Open Archives Object Reuse and Exchange (OAI-ORE). D'autres questions se posent sur la sélection des concepts, comme la nécessité

41. http://www.w3.org/DesignIssues/LinkedData.html
42. https://pro.europeana.eu/resources/standardization-tools/edm-documentation

Tableau 1. Caractéristiques principales des modèles basés sur FRBR et de leurs implémentations

	FRBRcore	**RDA**	**FRBRoo**	**FRBRer**	**BIBFRAME**	**LD4L / bibliotek-o**
Année	2005	2007	2008	2009	2012	2014 / 2017
Créateur	I. Danis, R. Newmann	RDA Steering Committee	IFLA	IFLA	LC	LD4L Labs
Type	Graphe	Vocabulaire	Orienté Objet	Entité Relation	Graphe	Graphe
Vocabulaires (hors base)	*core*	*rda, rdakit* FOAF, RE	*frbroo* FOAF, RE	*frbrer* FOAF, RE	*bibframe*	*bib*, bibframe PROV, OA, FOAF, datetime, RDAu schema, VIVO
Points forts	• Extensible • Première tentative	• Règles de catalogage	• FRBR + CRM • Méthodes pour des événements	• Respect strict de FRBR	• Distance par rapport à FRBR	• Version modifiée de BIBFRAME
Points faibles	• Pas d'attributs, que des classes • Incomplet	• Pas de scénarios	• Complexité	• Pas de hiérarchie • Interprétation fermée de FRBR	• Difficultés d' interopérabilité avec autres modèles	• « Proof of concept »
Exemples d' utilisation	B3Kat, HBZ	SUDOC, RLUK, UCL Cervantes	DLF et NUKAT, MUSES DOREMUS	datos.bne.es, (LIBRIS)	HCN, DNB Illinois, UCL, Naples	Cartographic Materials, ArtFrame

d'avoir des propriétés inverses, qui ne sont pas toujours disponibles, afin de faciliter l'écriture de requêtes et la navigation entre ressources. Par exemple, la *British Library* a complété la propriété *dcterms:creator* avec une propriété inverse *blt:hasCreated*, qui permet de lister plus simplement les contributions d'un agent.

Dans le web des données, chaque objet doit être représenté par une URI. La réutilisation d'URI (définies par d'autres sources) est une alternative à la création de ses propres URI, mais il n'y a pas forcément de garantie que la source externe conserve ses URI sur le long terme, et la définition externe peut ne pas inclure toutes les ressources descriptives nécessaires. Les bibliothèques choisissent donc en majorité de construire leurs propres URI. Une URI peut être transparente (i.e., avec une signification) ou opaque (i.e, un identifiant numérique). Dans le premier cas, l'URI est compréhensible par des humains, mais cela pose des difficultés par rapport au multi-linguisme. La fabrication d'URI transparentes reste un défi. Bien que la *British Library* ait défini des motifs de construction d'URI [43], celles-ci n'ont aucune garantie d'unicité (Deliot, 2014). Les URI opaques sont plus faciles à générer, mais ne sont pas explicites.

Enfin, les bibliothèques possèdent également de nombreux thésaurus (alias d'auteurs, catégories thématiques, etc.), et certains sont disponibles sur le web des données [44]. Par exemple, la *Library of Congress* a transformé ses vocabulaires contrôlés (Harper, Tillett, 2007). Comme les notices bibliographiques incluent de nombreux concepts répertoriés dans différents thésaurus, il est nécessaire de vérifier leur disponibilité sur le web des données afin d'anticiper leur exploitation pour l'étape de migration des données MARC.

Les recommandations quant à la modélisation sont d'éviter la prolifération de modèles et vocabulaires. Toutefois, même la réutilisation de modèles et vocabulaires nécessite une analyse minutieuse des concepts existants par rapport aux données à modéliser, afin de s'assurer que ces dernières puissent être correctement transformées vers le nouveau modèle. Lorsqu'un modèle a été sélectionné et que les choix de modélisation ont été décidés, l'étape suivante consiste à migrer les données du catalogue MARC vers le modèle sémantique.

3. Migration des données bibliographiques

Lorsque le modèle cible a été choisi, l'étape suivante de migration ou FRBRisation, consiste à transformer les données du catalogue MARC vers ce modèle (Aalberg, 2006). Rappelons qu'une notice MARC représente l'ensemble des informations d'une œuvre, i.e., elle peut décrire le travail intellectuel (e.g., titre, date de création), l'une de ses éditions (e.g., langue, titre traduit), les informations physiques (e.g., type de support physique) ou encore les personnes ayant contribué (e.g., créatrice, traductrice).

43. http://www.bl.uk/bibliographic/pdfs/british_library_uri_patterns.pdf
44. http://id.loc.gov/index.html

La FRBRisation doit d'abord interpréter les champs de la notice MARC pour générer des entités respectant les classes du modèle cible. Pour cela, des fonctions de transformation (correspondances) sont appliquées pour façonner un ou plusieurs champs MARC en une ou plusieurs entités du modèle cible. Dans l'exemple de la figure 1, la notice contient entre autres un champ 100 (personne principale), décomposée avec des sous-champs $a (nom de famille ou surnom) et $d (dates de naissance et/ou de décès). La migration du champ *100$a* produit une entité de type Agent avec une propriété nominative dont la valeur vaut *F'Murr*, et le champ *100$d* génère idéalement pour cette nouvelle entité deux propriétés distinctes. Les motifs bibliographiques sont des ensembles de relations entre entités, avec une signification particulière dans le domaine (e.g., adaptation, traduction). Elles constituent l'une des richesses des notices MARC, mais leur interprétation, qui requiert des informations de différents champs voire de différentes notices, est rendue complexe par les nombreuses règles implicites ou explicites utilisées pour les cataloguer. Enfin, l'interprétation génère de multiples entités équivalentes en plusieurs exemplaires (surtout pour les agents, œuvres et lieux), puisqu'une notice MARC contient l'ensemble des informations d'une œuvre. Un processus de déduplication finalise donc la migration.

Les solutions de FRBRisation les plus récentes se concentrent sur l'amélioration de la qualité de la migration, notamment pendant l'interprétation (Decourselle *et al.*, 2015). Cette section présente les travaux qui exploitent les données ouvertes et liées, que ce soit pour l'interprétation de champs simples et de motifs bibliographiques, ou pour la déduplication.

3.1. Interprétation des champs MARC

La transformation d'une notice MARC est une tâche difficile à cause de l'interprétation de certains champs et de leur valeur. En effet, des informations importantes peuvent être manquantes ou des pratiques locales de catalogage peuvent être utilisées. Par exemple, pour une bande-dessinée, la scénariste apparait avec la responsabilité principale (champ 700) en tant qu'auteure (sous-champ *$4* avec la valeur *070*), mais la dessinatrice, qui est stockée comme responsable secondaire (champ 702), ne possède pas de sous-champ *$4* pour identifier son rôle. De la même manière, la catégorie d'un livre peut contenir la valeur *r*, un code interne à la bibliothèque attribué aux romans. Dans de telles situations, il est difficile de transformer ces données correctement. Des outils génériques d'intégration de données peuvent être utilisés, par exemple Spoon-Pentaho Data Integration (Hallo *et al.*, 2014), ou des outils développés spécifiquement pour le contexte bibliographique, comme FRBR-ML (Aalberg, 2006), Metafacture [45], d:swarm [46] ou Syrtis [47]. En complément, les données ouvertes et liées peuvent aider à lever une partie des ambiguïtés de certains champs MARC.

45. https://github.com/metafacture/metafacture-core
46. https://github.com/dswarm/dswarm-documentation/wiki
47. https://www.progilone.fr/fr_FR/culture/

Une étape préliminaire à l'interprétation concerne le nettoyage des notices MARC. Par exemple, la *Bristish Library* a constaté que ses données d'origine n'étaient pas aussi cohérentes et standardisées qu'elle le pensait suite à la migration de son catalogue (Deliot, 2014). Vérifier la qualité des notices MARC (e.g., champs normés selon un thésaurus, format des titres) et nettoyer les champs incohérents via des sources externes peut améliorer la qualité d'une prochaine migration. Le benchmark BIB-R, qui permet d'évaluer les outils de migration bibliographique, prend en compte ce prérequis dans certains de ses tests (Aalberg *et al*., 2018). Les expérimentations réalisées avec BIB-R confirment le fait que les outils minimisent ce problème de qualité des catalogues MARC.

L'un des premiers outils permettant la FRBRisation est la plateforme libre eXtensible Catalog (XC). Elle offre plusieurs modules pour gérer un catalogue, dont le *Metadata Services Toolkit* qui est en charge de transformer les notices MARC en un modèle interne (XC schema) basé sur FRBR (Bowen, 2010). XC s'est principalement concentré sur la génération d'URI qui sont nécessaires pour publier des données liées. Chaque entité peut donc être référencée et avoir ses propres relations vers d'autres entités, y compris des référentiels externes. Le modèle interne de XC réutilise les concepts de RDA, *Dublin Core* et les *Library of Congress Subject Headings* (LCSH) pour présenter un catalogue « sémantique ». Dans les notices MARC, les catégories décrivant l'œuvre (e.g., poésie, médical) peuvent provenir d'un référentiel ou être librement saisies. Avec l'utilisation des LCSH, XC résout le conflit de libellés pour catégoriser toutes les œuvres d'un catalogue selon le même référentiel. L'algorithme de XC essaie d'aligner les valeurs des champs MARC aux LCSH, mais une intervention manuelle est requise pour valider ou corriger les correspondances *via* une interface graphique.

Un autre outil qui exploite les données liées est FRBR-ML (Takhirov *et al*., 2012). Il inclut un modèle hybride entre MARC et FRBR, en se basant sur l'hypothèse qu'une majorité de notices a une structure simple (une seule édition et une seule manifestation pour une œuvre). Ce modèle intègre des concepts du *Linked Open Data* (contraintes OWL-DL, vocabulaire FRBRer). Son format repose sur la syntaxe XML afin de favoriser les conversions entre différents formats (RDF notamment). La FRBRisation inclut un processus de correction des notices ambigües au moyen de trois stratégies, en particulier pour les champs 700 à 740 (« *added entries* »). Les deux premières stratégies consistent à repérer dans le catalogue et dans des catalogues externes des notices dans lesquelles l'ambiguïté est levée (e.g., même personne, mais dont le rôle est précisé). La dernière stratégie interroge les bases de connaissances DBpedia, Freebase et OpenCyc [48] afin de détecter l'entité correspondante *via* ses libellés alternatifs (alias). Les résultats de chacune des trois bases sont triés par pertinence, et l'entité avec le meilleur classement est considérée comme l'entité recherchée : son type (e.g., écrivain, dessinateur) est alors utilisé pour préciser le rôle manquant entre la personne et l'œuvre.

48. Les travaux sur Freebase et OpenCyc sont aujourd'hui arrêtés.

Des outils plus génériques peuvent également s'appliquer au contexte bibliographique. Par exemple, OpenRefine [49] permet entre autre de nettoyer des jeux de données, de les transformer et de les lier à d'autres sources. La bibliothèque de l'université du Nevada a montré les possibilités offertes par ces outils génériques (Lampert, Southwick, 2013). OpenRefine a été utilisé pendant les différentes phases de la migration : extraction, nettoyage des données (e.g., séparation du nom de l'auteur de ses dates), construction d'URI (selon des règles ou récupérées à partir des thésaurus de la LC), et implémentation des mappings vers un modèle basé sur FRBR.

Un autre outil utilisé dans la communauté bibliographique est X3ML [50]. Il est dédié d'une manière générale aux collections de données sur l'héritage culturel. L'article de Marketakis *et al.* rapporte les expériences de transformations pour de nombreux projets (Marketakis *et al.*, 2017). Le British Museum et le Rijkmuseum ont transformé leurs données hétérogènes vers le modèle CIDOC CRM, puis les ont intégrées dans l'environnement collaboratif *ResearchSpace* [51]. Dans le cadre du projet ARIADNE [52], des collections archéologiques, en particulier sur des pièces de monnaie romaines issues de différentes institutions, ont été migrées également vers CIDOC CRM avec X3ML.

L'interprétation des champs MARC reste un défi ouvert. Les approches qui exploitent les données ouvertes et liées ont montré une amélioration de la qualité lors de la transformation de certains champs MARC, mais celles-ci ne peuvent résoudre tous les problèmes d'interprétation rencontrés dans les notices. Lorsqu'un champ est incorrectement interprété, l'impact se limite généralement à la perte d'une seule information par champ (e.g., une propriété). À l'inverse, une interprétation erronée des motifs bibliographiques peut causer une perte d'informations plus importante.

3.2. Interprétation des motifs bibliographiques

Une grande difficulté pendant l'interprétation des notices MARC concerne la détection des motifs bibliographiques (Aalberg, Žumer, 2013). Ces motifs peuvent être vus comme des sous-graphes caractéristiques dans un modèle sémantique. Le motif le plus simple se compose d'une seule œuvre, d'une seule expression, d'une seule manifestation et d'une personne auteure de l'œuvre. En complément, on distingue quatre catégories de motifs bibliographiques plus complexes (Riva, 2004) :

– Les **augmentations** spécifient un contenu additionnel à une œuvre, mais dont le degré d'importance est faible (e.g., préface ou illustration de couverture). Par exemple, une notice [53] décrivant les *Essais* de Montaigne peut contenir un champ 200$a avec la valeur « *Essais... Chronologie et introduction par Alexandre Micha* ». L'augmentation,

49. http://openrefine.org/
50. https://github.com/isl/x3ml
51. http://www.researchspace.org/
52. http://www.ariadne-infrastructure.eu/
53. http://catalogue.bnf.fr/ark:/12148/cb331038613.unimarc

qui concerne ici la rédaction d'un texte introductif à l'œuvre, doit être dans ce cas déduite à partir de cette valeur. Les différents niveaux de titre sont aussi couramment utilisés pour indiquer la présence d'une augmentation;

– Les **dérivations** sont des œuvres issues d'une modification d'une autre œuvre (e.g., traduction, adaptation, révision). Par exemple, une notice [54] décrivant les *Essais* à la *Bristish Library* possède les métadonnées *245$a Essays* et *245$c Michel de Montaigne ; translated by M.A. Screech*. De plus, on y trouve le champ 500$a avec la note « *Translated from the French* » et un titre additionnel « *Michel De Montaigne ESSAIS* ». Ces multiples indices permettent de détecter une dérivation, mais ne sont pas toujours présents et restent difficilement interprétables automatiquement;

– Les **agrégations** représentent une relation d'ensemble avec ses composants. L'œuvre d'ensemble est parfois aussi importante que les parties qui la composent (e.g., *le seigneur des anneaux*). Les livres composés de différentes nouvelles (écrites par des auteurs différents) ou les pistes musicales dans une compilation sont d'autres exemples d'agrégations. Si nous illustrons ce motif sur les *Essais*, une compilation d'extraits existe et sa notice [55] s'intitule « *Le meilleur des Essais* ». Ce titre est complété par une mention « *textes choisis et présentés par Claude Pinganaud* » (champ 200$f), et on retrouve également « *Les Essais* » en titre uniforme (champ 500) et en variante du titre (champ 517). Dans ce cas, plusieurs champs permettent l'identification d'une agrégation;

– Les **œuvres complémentaires** modélisent des œuvres de même importance. Elles incluent les séquences (e.g., une œuvre est la suite d'une autre) ainsi que les œuvres d'accompagnement, indissociables d'une autre œuvre (e.g., un manuel d'exercice et le livret avec corrigés). Les *Essais* ont été édités de différentes façons, y compris en plusieurs volumes. Nous pouvons donc trouver des notices décrivant différentes parties de cette œuvre. Une notice [56] contient le champ 200$a « *Essais de Michel de Montaigne* » suivi du champ 200$h « *Livre premier* » (indiquant un numéro de partie). Une autre notice [57] possède le même champ 200$a mais un champ 200$h évalué à « *Livre second* ». Le titre identique et la présence d'une métadonnée sur la numérotation facilite dans ce cas l'interprétation d'une œuvre complémentaire.

La détection de ces motifs nécessite généralement d'interpréter correctement plusieurs champs. Il est donc fréquent de ne découvrir qu'une partie du motif (e.g., détection d'une traductrice, mais pas de la version traduite), ce qui empêche ou restreint la FRBRisation du motif. Par exemple, dans le cas d'une traduction mal interprétée, nous perdons le fait qu'il existe une traduction dans une autre langue, le titre traduit, les informations sur la traductrice, etc. Certains champs et valeurs associées autorisent une transformation automatisée tout en garantissant une bonne qualité par rapport à la détection des motifs (e.g., la publication des *Essais* en plusieurs volumes). Mais dans

54. http://primocat.bl.uk/F/?func=direct&doc_number=016488535&format=001
55. http://catalogue.bnf.fr/ark:/12148/cb38919858x
56. http://catalogue.bnf.fr/ark:/12148/cb420586293
57. http://catalogue.bnf.fr/ark:/12148/cb420580988

une majorité de cas, les champs sont complexes à interpréter même avec des outils de traitement automatique du langage (e.g., la traduction des *Essais* à déduire de la description textuelle en note ou qui accompagne le titre). Le benchmark BIB-R propose des jeux de données permettant d'évaluer un outil de FRBRisation pour chaque motif bibliographique et rend compte de ces nombreuses difficultés, surtout lorsque l'on considère les pratiques de catalogage internes qui diffèrent d'une institution à l'autre (Aalberg *et al.*, 2018).

La FRBRisation de WorldCat, le catalogue unifié le plus important avec plus de 2 milliards d'œuvres à ce jour, commence au début des années 2000 (Hickey, O'Neill, 2005). Les auteurs de ces travaux utilisent l'algorithme *WorkSet* pour regrouper les *Works*, qui se basent sur une combinaison d'attributs comme le titre et le créateur normalisés par le référentiel *Library of Congress Name Authority File* (LCNAF) [58]. Ils identifient majoritairement des motifs simples (une seule *Manifestation* pour un *Work*), au total 36 millions de cas. Mais ils s'intéressent également aux augmentations, révisions, agrégations et traductions, qui comptent collectivement environ 1,5 million d'œuvres. Des œuvres intensivement étudiées dans la littérature (e.g., *The Expedition of Humphry Clinker*, par Tobias Smollett) sont sélectionnées pour évaluer la détection de chacun des quatre types de motif et décrire les obstacles quant à leur interprétation.

Les travaux de He *et al.* utilisent les données liées pour FRBRiser les mangas (He *et al.*, 2013). En effet, ces derniers ont une organisation complexe : un titre de manga regroupe plusieurs histoires, et une histoire se découpe en épisodes. Un magazine inclut des épisodes ou histoires de différents mangas (agrégation), et les histoires les plus populaires sont rééditées sous forme de monographes (agrégation et augmentation), qui sont exportés à l'étranger (dérivation, avec titres très hétérogènes). L'identification automatique des œuvres est réalisée en regroupant celles-ci par langue, puis en exploitant DBpedia (langage et liens *owl:sameAs*) pour détecter les traductions. La version japonaise de DBPedia permet de détecter les magazines dans lesquels une œuvre a été initialement publiée (propriété *dbpprop-ja:keisaishi*). Seuls 20 % des mangas sont transformés correctement (absence d'entité DBpedia correspondante pour la majorité des 80 % restants), mais les auteurs montrent aussi que la détection de l'œuvre, de ses traductions et de ses agrégations n'est pas possible avec le catalogue original.

À notre connaissance, il n'y a pas d'autre approche qui exploite les données liées à ce niveau [59]. Pourtant, ces dernières offrent un potentiel non négligeable d'amélioration de la qualité pour transformer les motifs bibliographiques. Par exemple, DBpedia contient les relations de précédence ou de suite entre plusieurs œuvres, ce qui permet d'identifier une agrégation. Dans le catalogue FRBRisé de la Bibliothèque Nationale de France (BNF), les traductions sont fortement présentes et peuvent confirmer un motif de dérivation. Vu l'hétérogénéité des sources externes, la difficulté est d'ap-

58. L'algorithme WorkSet est un peu plus complexe : un système de priorité pour les différents champs de titrage ou d'auteur, une normalisation des valeurs, ou des variantes qui exploitent des champs additionnels.

59. Notons que l'outil FRBR-ML, décrit en section 3.1, résout certains problèmes d'interprétation pour les motifs.

prendre à reconnaitre les correspondances nécessaires pour identifier automatiquement les concepts importants d'un motif bibliographique. Quand l'interprétation des notices est terminée, l'ensemble d'entités produites contient des entités redondantes (e.g., auteur décrit dans chacune de ses œuvres), qu'il est donc nécessaire de nettoyer.

3.3. Déduplication des entités

La détection d'entité équivalentes dans une collection de données (déduplication) ou entre plusieurs collections (alignement d'entités, résolution d'entités ou liage d'enregistrement) est un domaine de recherche étudié depuis plusieurs décennies dans la communauté gestion de données (Newcombe *et al.*, 1959 ; Fellegi, Sunter, 1969 ; Christen, 2012a). Dans le contexte bibliographique, la déduplication est également un problème connu. Avant l'intérêt pour les catalogues sémantiques, le défi consistait en la détection de notices équivalentes. Sitas *et al.* ont comparé dix algorithmes de détection de notices utilisés par des institutions telles que l'*Online Computer Library Center* (OCLC), les bibliothèques académiques de Grèce, le réseau des bibliothèques académiques d'Israël ou encore des universités états-uniennes ou anglaises (Sitas, Kapidakis, 2008). Cette étude insiste sur la difficulté à produire une solution générique, puisque chaque institution possède ses propres pratiques. Dans le cas des catalogues sémantiques, le problème de la déduplication devient plus général, et se rapproche donc de celui de la communauté gestion de données. Dans les exemples mentionnés en Section 3.2 sur les *Essais* de Montaigne, l'interprétation de chaque notice produirait (idéalement) un agent *Michel de Montaigne* à partir des champs de responsabilité ou des titres, et une œuvre *Essais* à partir des champs titres ou notes. Si plusieurs de ces notices appartiennent au même catalogue à FRBRiser, des entités équivalentes pour l'agent ou pour l'œuvre apparaissent dans le résultat de l'interprétation et il est nécessaire de les dédupliquer. Lorsqu'un identifiant (interne ou vers un référentiel) n'est pas disponible, la comparaison des entités se fait généralement en utilisant un sous-ensemble des attributs de chaque paire d'entités. Lors d'une étape dite de *blocking*, cette comparaison doit être rapide (quelques attributs utilisés) puisque l'objectif est de créer des groupes d'entités. Lors de l'alignement, les paires d'entités d'un même groupe sont comparées plus finement, par exemple en utilisant des algorithmes plus sophistiqués ou un plus grand nombre d'attributs. Les principaux verrous scientifiques concernent le regroupement des entités (*blocking*), la comparaison d'entités (alignement) et la fusion d'entités détectées comme équivalentes.

Le processus de déduplication peut s'avérer très coûteux s'il est réalisé avec un produit cartésien, i.e., en comparant toutes les entités (ou notices) deux à deux. Le *blocking* permet donc de regrouper des sous-ensembles d'entités qui seront ensuite comparées plus en détail (Christen, 2012b). Comme le *blocking* se limite à des algorithmes basiques (e.g., égalité sur l'année de création ou sur le titre de l'œuvre), des entités équivalentes peuvent être manquées, notamment à cause des nombreux alias (e.g., de personnes, de lieux) ou à cause du multi-linguisme (e.g., pour le titre de l'œuvre). Par exemple, le système MELVYL, dont l'objectif est d'unifier les catalogues des campus californiens (Coyle, 1992), dispose d'un *blocking* qui s'effectue

en regroupant les notices possédant un même numéro ISBN, un même numéro LCCN (Library of Congress Control Number), ou un titre avec les 25 premiers caractères identiques. Cet algorithme de *blocking* montre ses limites dans le cas où les titres diffèrent sur leurs premiers caractères. Les jeux de données ouverts et liés comme VIAF ou DBpedia offrent une liste de noms alternatifs ou des titres dans plusieurs langues, et peuvent donc être utilisés pour regrouper des entités équivalentes qui ne satisfont pas aux algorithmes basiques, ce qui peut améliorer la qualité de la déduplication.

Dans le contexte bibliographique, l'alignement des entités est tout d'abord réalisé à la volée, c'est à dire que les informations sont encore stockées en MARC mais qu'une déduplication est réalisée suite à une requête utilisateur. C'est le cas du système PRIMO, qui regroupe des œuvres selon les principes FRBR malgré un stockage physique sous forme de notices (Sadeh, 2007). Peu de détails sont fournis sur la méthode utilisée pour la déduplication : PRIMO se base sur « *les types d'entités FRBR ainsi que sur des améliorations résultant du feedback des bibliothécaires et des partenaires de développement* ». Dans le catalogue public OpenLibrary [60], qui contient plus de 20 millions de notices bibliographiques et 6 millions d'auteurs, l'ajout de nouvelles ressources est un processus collaboratif. Comme tout le monde peut contribuer, certaines œuvres se retrouvent plusieurs fois sur le site. La déduplication est réalisée manuellement pour nettoyer le catalogue, mais une recherche populaire sur le site (e.g., « Tolkien ») illustre les limitations d'un traitement manuel. Les travaux de Freire *et al.* étudient l'identification et la déduplication de *Works* FRBR (Freire *et al*., 2007). L'expérimentation porte sur le catalogue PORBASE de la bibliothèque nationale portugaise, mais des tests préliminaires sont réalisés sur celui d'un éditeur portugais (Porto Editoria), dont les notices sont un sous-ensemble de PORBASE. L'algorithme est une moyenne pondérée de deux valeurs, calculées avec Jaro-Winkler sur les auteurs et avec Jaro-Winkler-TFIDF sur les titres. La décision de considérer deux *Works* comme équivalents est prise selon une valeur seuil. Les résultats montrent entre autre qu'il y a en moyenne trois *Works* par notice après déduplication. Les auteurs s'interrogent également sur l'applicabilité de cet algorithme pour dédupliquer les *Expressions*. Le système MELVYL implémente deux algorithmes d'alignment (Coyle, 1992). L'algorithme simplifié n'utilise que le numéro LCCN, le titre et la date. En cas d'échec de cette version simple, l'algorithme complet et pondéré compare les valeurs d'une vingtaine de champs (titre, auteurs, dates, édition, pagination, etc.) selon différentes heuristiques. Par exemple, la valeur maximale du poids est utilisée si les dates des deux notices sont identiques, alors qu'une valeur moindre est accordée si elles diffèrent de une ou deux années. Cet article souligne déjà le fait que des identifiants externes avec la même valeur (ISBN et LCCN) nécessite tout de même une vérification supplémentaire sur d'autres champs (titre et date dans ce cas). Mais 90 % du catalogue est tout de même dédupliqué grâce à l'algorithme simplifié, d'où l'importance des identifiants de référence. L'article de Hammerton *et al.* s'intéresse à la déduplication des notices sur la plateforme Mendeley (Hammerton *et al*., 2012). Comme chaque utilisateur de cette plateforme peut référencer, importer, et annoter des articles dans son espace, il

60. http://openlibrary.org/

est nécessaire de détecter les articles doublons (importés par différents utilisateurs) pour améliorer la qualité de la recherche et des recommandations. Toutefois, les auteurs constatent eux aussi qu'un même identifiant arXiv ou PubMed ne garantit pas l'équivalence entre deux notices. Bien que la déduplication utilisée dans les systèmes MELVYL et Mendeley porte sur les notices, il est intéressant de noter que leurs auteurs privilégient l'exploitation de sources externes comme solution. Dans le contexte bibliographique, des identifiants externes (e.g., ISBN, identifiants vers les notices des bibliothèques nationales) sont parfois disponibles dans les notices d'origine, et généralement conservés pendant l'interprétation. Ils peuvent donc être exploités pour faciliter la déduplication. Les identifiants vers des entités du LOD (e.g., DBpedia, Wikidata, VIAF), de plus en plus souvent disponibles (voir section 4.2), peuvent aussi faciliter la déduplication de certaines entités (*Work* et *Agent* essentiellement), en particulier pour résister au problème de multi-linguisme.

Lorsque deux entités sont détectées comme équivalentes, il est nécessaire de les fusionner (Dong *et al*., 2014). Bien que peu détaillée dans la littérature, la fusion de notices a été un objectif crucial dans le cadre de l'unification de catalogues. Par exemple, dans le système MELVYL (Coyle, 1992), les règles de fusion se limitent souvent à normaliser des chaînes de caractères dont la ponctuation varie (e.g., présence ou non d'un trait d'union dans un titre). En présence de plusieurs valeurs pour une même propriété, les référentiels exposés sous forme de données liées et ouvertes peuvent aider à uniformiser la forme des titres, des concepts, etc. Par exemple, un référentiel pourrait proposer les titres sous la forme *Génie des alpages, (Le)* plutôt que *Le génie des alpages*.

La déduplication reste nécessaire lorsqu'un catalogue est disponible sous forme de données liées et ouvertes. En effet la réception – parfois quotidienne – de nouvelles notices implique la détection des entités équivalentes entre le catalogue et les nouvelles notices. Ce processus doit être performant, particulièrement si le catalogue est composé de millions d'entités. Après l'interprétation et la déduplication, le catalogue FRBRisé est généralement vérifié et testé par des experts. Cependant, il reste encore isolé et difficilement accessible sans la création de liens vers d'autres sources de données.

4. Liage et enrichissement

Quand le catalogue est transformé, il est désormais fréquent de le lier à d'autres sources de données, qu'il s'agisse de référentiels et catalogues publiés sous forme de données ouvertes et liées ou de bases de connaissances qui favorisent la sérendipité. En effet, l'intérêt qu'un lecteur porte à une œuvre peut concerner non seulement l'œuvre elle-même ou d'autres œuvres en lien (*e.g.,* suite, adaptation, autres œuvres du même auteur, étude sur l'œuvre), mais aussi la vie de l'auteur, son environnement lors de la création de l'œuvre ou des faits ou événements en lien avec l'œuvre telles que des lectures publiques ou spectacles. Il existe différentes méthodes pour lier un catalogue à d'autres sources (Cole *et al*., 2013). Pour communiquer avec une source externe, il

est nécessaire d'établir les correspondances entre le modèle du catalogue et celui de la source externe (alignement d'ontologies). Nous distinguons la création de lien vers une source externe, qui suppose d'interroger la source distante pour récupérer d'autres informations, et l'enrichissement, qui ajoute ou fusionne directement des informations de sources externes au catalogue. La section passe également en revue les catalogues disponibles sous forme de données ouvertes et liées.

4.1. Alignement d'ontologies

Le catalogue sémantique se conforme à un modèle donné (voir section 2), qui diffère généralement des modèles des sources de données externes (e.g., DBpedia possède sa propre ontologie). Par exemple, dans WorldCat, un titre d'œuvre est représenté par la propriété `schema:name` alors qu'il est stocké avec les propriétés `rdfs:label` ou `dct:title` à la *British Library*. De la même manière, un nom d'auteur chez WorldCat se trouve dans la propriété `rdfs:label` tandis que la *British Library* utilise `foaf:name` et `foaf:familyName`. L'alignement concerne aussi les classes, par exemple pour établir la correspondance entre l'*Instance* de BIBFRAME et les classes *Expression* et *Manifestation* de FRBRer. L'alignement d'ontologies est un domaine de recherche prolifique depuis de nombreuses années qui vise à étudier la détection de concepts issus de différentes ontologies mais possédant une relation sémantique comme l'équivalence ou la subsomption (Euzenat, Shvaiko, 2013). Ce processus peut servir à la fois pour la déduplication (e.g., alignement avec un référentiel ou une source LOD afin de détecter les doublons d'un catalogue), pour le liage ou l'enrichissement (e.g., détection de classes ou propriétés équivalentes afin de relier ou fusionner des entités) et pour compenser les problèmes de multiplication des modèles/vocabulaires de description de métadonnées. L'alignement d'ontologies est supporté par la compétition annuelle *Ontology Alignment Evaluation Initiative* (OAEI) [61] depuis 2004. Entre quinze et vingt outils concourent chaque année à OAEI, et nous renvoyons donc vers une étude récente pour une liste exhaustive des outils disponibles ainsi que différentes statistiques sur leur évolution ou la qualité obtenue lors de leur participation à OAEI (Otero-Cerdeira *et al.*, 2015). Un défi « *Library* » a été proposé à OAEI entre 2012 et 2014 pour aligner deux thésaurus SKOS qui décrivent les thèmes en sciences sociales et économie. Dans notre contexte, les ontologies et vocabulaires bibliographiques sont souvent utilisés par plusieurs institutions et (relativement) stables dans le temps. Il serait donc intéressant de créer et de partager les correspondances, avec un standard comme *Alignment format* [62] ou son extension *Expressive and Declarative Ontology Alignment Language* (EDOAL) [63]. Quelques travaux proposent des alignements (partiels) entre BIBFRAME et EDM (Zapounidou *et al.*, 2014) ou entre FRBRer et BIBFRAME (Zapounidou *et al.*, 2017). Enfin, la réuti-

61. http://oaei.ontologymatching.org/
62. http://alignapi.gforge.inria.fr/format.html
63. http://ns.inria.org/edoal/1.0/

lisation de vocabulaires non spécifiques au domaine (FOAF, DC, schema.org, etc.) facilite l'alignement entre un catalogue et une source externe.

4.2. Création de liens

Le liage des données (*entity linking* ou *data interlinking*) est un domaine très étudié, en particulier pour les personnes et les œuvres (Shen *et al.*, 2015). Contrairement au *Named Entity Recognition*, qui s'attache à identifier une mention d'entité et son type, le liage d'entités permet une meilleure désambiguïsation en reliant cette mention à l'entité correspondante sur une base de connaissances ou référentiel. Un catalogue sémantique est aujourd'hui peu intéressant s'il n'est pas relié à d'autres sources de données. L'avantage le plus évident est la possibilité de fournir des informations supplémentaires aux utilisateurs (e.g., retrouver toutes les œuvres d'une personne ou naviguer par sérendipité). Des référentiels comme VIAF [2] permettent d'identifier sans ambiguïté une personne, de normaliser des valeurs (e.g., nom), ou encore de détecter facilement des liens vers d'autres sources (puisque ces dernières essaient aussi de se lier aux référentiels). Pour l'exemple du *génie des alpages*, il existe un mémoire [64] (*Le Monde insolite du Génie des Alpages*, datant de 1980, publié à l'ENSSIB, l'école nationale supérieure des sciences de l'information et des bibliothèques) dont le sujet est l'œuvre principale de F'Murr. Il serait intéressant de relier ces deux ressources (avec une propriété `rda:subjectWork` par exemple), mais il n'existe à notre connaissance aucune source possédant cette relation, du moins jusqu'à la migration du catalogue de l'ENSSIB.

Les algorithmes de liage peuvent être génériques comme dans les outils FRED (Consoli, Recupero, 2015) et KARMA [65], ou développés spécifiquement pour une collection (e.g., en construisant des requêtes SPARQL). Dans cette partie, nous décrivons des applications d'outils génériques pour lier un catalogue bibliographique.

L'outil OpenRefine a déjà été mentionné pour l'interprétation des champs MARC (voir section 3.1). Cet outil permet aussi de créer des liens vers d'autres sources de données [66], parmi lesquelles VIAF, VIVO, la terminologie LCSH, DBpedia et d'autres plus spécifiques pour des domaines comme le médical, la biologie ou la numismatique. L'université du Nevada relate ses expérimentations d'alignement entre ses collections et les termes du LCSH en utilisant OpenRefine (Lampert, Southwick, 2013).

La plateforme SILK [67] permet de transformer, publier, et lier ses données sous forme de LOD (Volz *et al.*, 2009). Le module de liage utilise une dizaine de mesures de similarité (e.g., Jaro, q-grams), qui peuvent être combinées par cinq opérateurs (moyenne, produit, distance euclidienne, min et max). Une quinzaine de fonctions de

64. http://www.enssib.fr/bibliotheque-numerique/notices/63038-monde-insolite-du-genie-des-alpages-de-f-murr

65. https://usc-isi-i2.github.io/karma/

66. https://github.com/OpenRefine/OpenRefine/wiki/Reconcilable-Data-Sources

67. http://silkframework.org/

transformation permettent d'améliorer la détection d'entités, par exemple en concaténant deux valeurs ou en retirant des caractères spéciaux. Une valeur seuil permet de filtrer la ou les entité(s) équivalentes parmi toutes celles candidates. L'outil SILK a notamment été utilisé pour aligner une collection d'articles scientifiques FRBRisés vers les jeux de données VIAF, Geonames et DBLP (Hladka *et al*., 2012). De même, il a permis de lier la collection de notices MARC de la « *Electrical Enginnering Library* » à OpenLibrary, Europeana et LCSH (Hallo *et al*., 2014).

L'outil FRBRpedia propose de transformer un article de type culturel disponible sur Amazon en une représentation FRBRisée et liée à d'autres sources (Duchateau *et al*., 2011). La transformation selon le modèle FRBR dépend du succès à aligner le produit vers l'entité WorldCat correspondante. Un algorithme appliquant des mesures de similarité (sur les titres, les auteurs, et les catégories) couplé à un algorithme de tri des résultats permet de sélectionner l'entité DBpedia pour le produit. Une extension de cet outil aligne un catalogue bibliographique avec DBpedia (Takhirov *et al*., 2011), et la vérification manuelle des liens générés indique une qualité acceptable (80 % F-score). Pour des raisons de performance, une étape de *blocking* permet de ne récupérer que des entités candidates de DBpedia parmi les plus pertinentes, en combinant des attributs pré-définis comme le titre, la personne créatrice ou le type d'œuvre. L'alignement de 700 œuvres a pris au final une dizaine de minutes.

L'outil KARMA a été utilisé pour publier les 41 000 objets du musée *Smithsonian American Art* sous forme de LOD (Szekely *et al*., 2013). Les sources visées sont DBpedia, le *Getty Union List of Artist Names* (ULAN) et le jeu de données du Rijksmuseum afin d'y aligner les artistes du musée. L'algorithme utilise le nom des artistes avec des variantes, ainsi que les dates de naissance et décès, et calcule un score de similarité dans [0, 1] en exploitant à la fois des statistiques sur les années et Jaro-Winkler pour le nom. L'objectif est d'obtenir une précision élevée en prévision d'une vérification manuelle par les conservateurs. L'évaluation sur un échantillon de 535 artistes produit un degré de qualité élevé (96 % F-score, pour une précision de 99 %). Environ 5 000 liens vers des entités des trois sources ont été finalement proposés aux conservateurs.

Bien que les approches de liage d'entités abondent, de nombreux verrous spécifiques au domaine bibliographique subsistent, par exemple pour lier une Expression ou une Manifestation vers WorldCat, OpenLibrary, ou les catalogues FRBRisés des bibliothèques nationales. Les projets de catalogue LOD présentés dans la section suivante témoignent de ces difficultés.

4.3. Jeux de données LOD

De nombreuses institutions ont proposé leur catalogue sous forme de données ouvertes et liées. Nous décrivons dans cette section les principaux projets en insistant sur les méthodes de liage utilisées.

Le *Online Computer Library Center* (OCLC) est responsable de deux jeux de données LOD de première importance. Le catalogue WorldCat inclut de nombreux liens vers les bibliothèques nationales, des référentiels et d'autres sources du LOD (Teets, Goldner, 2013). Ce sont principalement les institutions qui ont ajouté les liens de leur catalogue national vers celui de WorldCat, mais il y a peu de détails techniques sur la création de liens vers les autres sources comme Wikipedia. La base *Virtual International Authority File* (VIAF) est un effort international qui fusionne les fichiers d'autorité (i.e., personnes, organisations) d'une soixantaine de bibliothèques. C'est aujourd'hui un référentiel incontournable auquel de nombreux jeux de données se lient.

Les grandes bibliothèques nationales ont été parmi les premières à s'intéresser au liage de leur catalogue (Hallo *et al.*, 2016). Pour atteindre cet objectif, une étape d'interprétation, voire de FRBRisation partielle, a été nécessaire, suivie d'une étape d'alignement. Ces étapes, qui sont souvent réalisées par un tiers (e.g., Logilab pour la Bibliothèque Nationale de France, Talis pour la *British Library*), ne sont pas forcément bien détaillées.

La *Library of Congress* (LC) a rapidement transformé et publié ses fichiers d'autorité comme le *LC Name Authority File* (LCNAF, mais plus communément appelé NACO) pour les agents (personnes et organisations), ou les *LC Subject Headings* (LCSH) pour une classification de descripteurs (Harper, Tillett, 2007). Ce vocabulaire contrôlé est également une source fréquente d'interconnexion dans le domaine des bibliothèques.

En Allemagne, la bibliothèque nationale (DNB) a transformé ses fichiers d'autorité dès 2010 (Hannemann, Kett, 2010), et son catalogue de notices est publié sur le LOD en 2012. Basé sur les principes FRBR de RDA, le fichier d'autorité unifié (GND) bénéficie dès le départ de liens vers Wikipedia, LCSH, VIAF, DBpedia et RAMEAU grâce à de précédents projets.

La *British National Bibliography* est une collection exposée sous forme de données liées et ouvertes depuis 2011 (Deliot, 2014). Elle est notamment reliée au LCSH et à VIAF. L'alignement automatique des entités se base sur la comparaison exacte entre libellés d'autorités et libellés des sources externes, mais aussi sur l'extraction d'informations pour les champs codés.

En 2013, le catalogue de la bibliothèque nationale espagnole a été FRBRisé puis lié à VIAF (liens présents dans les notices) et à Lexvo [68] pour des informations linguistiques *via* le champ *dcterms:language* (Vila-Suero *et al.*, 2013). Les 450 000 liens vers VIAF permettent de construire directement 130 000 liens *owl:sameAs* supplémentaires vers d'autres catalogues de bibliothèques nationales et vers DBPedia.

En France, la BNF a transformé une partie de son catalogue. Ce jeu de données, *data.bnf.fr*, a d'abord produit des liens vers VIAF en se basant sur des comparaisons exactes (Simon *et al.*, 2013). Par extension, ces liens vers VIAF ont permis de détecter,

68. http://www.lexvo.org/

en comparant approximativement les chaînes de caractères, de nouveaux liens vers d'autres sources, dont LCSH, DBpedia, DNB, Geonames (170 000 liens). Pour aligner des cas plus complexes, un algorithme d'apprentissage (*logistic regression*) détermine si une notice est représentative de l'œuvre, auquel cas les notices moins représentatives peuvent être corrigées.

En Finlande, la bibliothèque nationale a converti plus d'un million de notices MARC vers le modèle BIBFRAME (pour la séparation des *Works* et des *Instances*) et vers `schema.org` (Suominen, 2017). Cette solution a été pensée pour favoriser l'interopérabilité avec d'autres sources du web. Des liens sont principalement créés vers les fichiers d'autorité nationaux (personnes, lieux, concepts), les LCSH et Wikidata. Les Works et Instances ne sont donc pas encore reliés, mais les auteurs planifient de créer des liens vers VIAF ou WorldCat.

L'union des catalogues suédois (LIBRIS), propose un catalogue de sept millions de titres [69]. L'un des premiers objectifs était de relier les entités au sein d'un même catalogue, puis vers des sources externes (Malmsten, 2008). Les liens créés se basent sur des clés descriptives (e.g., auteur, année et titre). Le catalogue est lié à Wikipedia et DBpedia, et inclut des annotations grâce au vocabulaire `annotea` [70].

Le projet de catalogue sémantique des bibliothèques suisses, *linked.swissbib.ch*, s'attelle à transformer 21 millions de notices en entités (dans une version très simplifiée de FRBR, avec uniquement les classes *Person* et *BibliographicResource*), puis de les lier et les enrichir grâce au LOD (Bensmann *et al.*, 2017). Pour lier les auteurs à DBpedia et VIAF, les auteurs disposent de quatre propriétés : nom, prénom, date de naissance et date de décès (ces deux dernières étant indisponibles dans une majorité de cas). Les aspects performance sont importants dans le contexte de *linked.swissbib.ch* : un *blocking* est exécuté en utilisant la première lettre du nom de la personne afin de réduire le nombre de personnes à comparer. Les blocs sont limités à 200 000 personnes, quitte à créer plusieurs blocs pour la même lettre. Les auteurs ne précisent pas comment deux blocs portant sur la même lettre sont comparés. Au final, plus de 30 000 liens vers DBpedia et 20 000 liens vers VIAF sont établis. En vérifiant un sous-ensemble (100) des liens découverts selon plusieurs critères, les auteurs obtiennent une précision élevée quand trois propriétés sont prises en compte (nom, prénom et date de naissance).

Le projet *data.europeana.eu* agrège quant à lui 1 500 collections provenant de nombreux pays (Haslhofer, Isaac, 2011) pour un total de 52 millions d'œuvres. Elles sont reliées à VIAF, DBpedia, Geonames et GEMET [71] pour les sujets thématiques. Le liage est réalisé par un processus qui réutilise les données des collections initiales (e.g., alignement de la valeur de la propriété *dc:spatial* avec les labels de Geonames).

69. http://libris.kb.se/

70. https://www.w3.org/2001/Annotea/

71. http://www.eionet.europa.eu/gemet/en/themes/

Au *Leibniz Information Center for Economics* (ZBW), plusieurs projets de données ouvertes et liées ont été développés, notamment une plateforme sémantique d'articles scientifiques dans le domaine économique nommée EconStor [72] (Latif *et al.*, 2014). Son modèle de données est inspiré de DSpace [73], et les vocabulaires utilisés sont standards (RDFS, Dublin Core, FOAF) en plus de *Semantic Web for Research Communities* (SWRC) spécifique à la recherche. La construction des URI pour les ressources suit les recommandations de DSpace (i.e., spécification Handle [74]). Le catalogue est relié à deux thésaurus économiques, Lexvo pour les langues, le fichier d'autorité allemand GND, AGROVOC [75] pour l'alimentation et l'agriculture et DBpedia. L'alignement avec le GND est financé dans le cadre du projet national KoMoHe (2004-2007) pour établir manuellement des « correspondances croisées », en se focalisant sur des relations d'équivalence, de hiérarchie, et d'association entre termes (Mayr, Petras, 2008). Pour les thésaurus économiques, l'alignement a été réalisé dans le cadre de la campagne OAEI, et les outils offrant de bons résultats ont été par la suite utilisés pour la découverte de correspondances vers AGROVOC et DBpedia (Kempf, Neubert, 2016).

De nombreuses autres institutions (bibliothèques de grandes villes, musées) partagent leur expérience de publication de leurs collections sous forme de LOD, à travers des présentations comme celles des conférences *Semantic Web in Libraries* (SWIB). Quelques exemples : la collection EMBLEMATICA composée de textes de la Renaissance qui s'enrichit grâce à VIAF, WorldCat ou la classification multilingue IconClass (Han *et al.*, 2016) ou la collection de textes médiévaux espagnols interconnectée à une une douzaine d'autres sources LOD (Cruz, Testal, 2013).

Les expérimentations présentées dans des articles scientifiques décrivent généralement plus en détail les processus de liage. Les travaux de Candela *et al.* ont permis de migrer le catalogue de la *Biblioteca Virtual Miguel de Cervantes* (200 000 notices et 50 000 fiches d'autorité) et de l'exposer sous forme de données liées (Candela *et al.*, 2016). En particulier, le catalogue a été lié à DBpedia pour les œuvres (500 relations *owl:sameAs*) et à VIAF et ISNI pour les personnes (6 500 liens). La technique utilisée pour le liage d'entités n'est pas décrite. À la bibliothèque d'Oslo, une expérimentation de FRBRization et de liage a été réalisée en 2012 pour deux auteurs norvégiens populaires (Westrum *et al.*, 2012). La centaine d'œuvres identifiées (après nettoyage) a été manuellement liée à DBpedia, VIAF et au Projet Gutenberg. Cette expérimentation a montré l'importance d'identifier, *via* les sources externes, l'œuvre principale et un accès rapide au texte, sans nécessairement proposer à l'utilisateur les dizaines de versions disponibles. Un autre article relate la transformation d'une collection d'articles scientifiques de l'université d'économie de Prague en des données ouvertes et liées (Hladka *et al.*, 2012) et leur liage vers VIAF, Geonames et DBLP en utilisant

72. https://www.econstor.eu/
73. https://duraspace.org/dspace/
74. https://fr.wikipedia.org/wiki/Handle_System
75. http://aims.fao.org/fr/agrovoc

la plateforme SILK (Jentzsch *et al*., 2010). L'une des difficultés porte sur l'ambiguïté des noms de lieux géographiques. L'ajout du nom de l'éditeur a permis de résoudre ce problème. Les auteurs de la collection possèdent des noms parfois similaires, et une correction manuelle a été nécessaire. Enfin, l'étude précise que les jeux de données ouverts évoluent fréquemment, et qu'il est nécessaire de relancer le processus d'appariement régulièrement pour mettre à jour les liens. Dans un domaine proche, les travaux de Raimond *et al.* portent sur l'interconnexion de jeux de données musicaux (Raimond *et al*., 2008). Trois algorithmes d'appariement sont présentés : une version naïve, un second avec extension des littéraux, et enfin un troisième qui exploite le voisinage de l'œuvre (i.e., albums et pistes) à travers une mesure de similarité dans un graphe. L'une des expérimentations consiste à aligner des artistes et œuvres musicales entre une collection personnelle modélisée par l'ontologie Music Ontology (basée sur les spécifications de FRBR) et MusicBrainz, l'un des catalogues de métadonnées les plus importants au niveau musical. L'évaluation ne porte que sur une seule œuvre (populaire et reprise par de nombreux artistes), pour laquelle différentes modifications sont apportées (e.g., suppression de l'auteur, nom de l'album remplacé par une chaine de caractères aléatoire). Cette vingtaine de modifications représente des cas typiques d'erreur dans les métadonnées. Les auteurs soulignent les difficultés d'alignement lorsque des titres sont similaires ou identiques pour plusieurs artistes, mais l'algorithme est globalement résistant à la majorité des modifications.

Dans le domaine bibliographique, de nombreux catalogues sont aujourd'hui (partiellement) disponibles sous forme de LOD. Les solutions adoptées pour obtenir ce liage sont souvent manuelles, basées sur des liens ou identifiants déjà existants ou reposant sur une implémentation interne. Cette exposition soulève de nombreux problèmes de qualité (vérification manuelle, expertise en informatique et de bibliothécaire) et reste limitée à certains types d'entités (agents, œuvres, au détriment des concepts plus spécifiques du domaine comme les expressions ou manifestations de FRBR).

4.4. Enrichissement du catalogue

L'utilisation de données ouvertes et liées ouvre la perspective aux bibliothécaires de pouvoir enrichir leur catalogue et ainsi en améliorer la qualité, en d'autres termes, améliorer automatiquement la complétude et la cohérence des entités. S'il est plutôt conseillé de créer un lien vers une entité externe, notamment à cause de la problématique de mise à jour des informations, il peut tout de même s'avérer intéressant d'intégrer directement, soit pour des raisons de performances (pas d'accès distant à l'information), soit pour des informations intemporelles (e.g., date de naissance). Il s'agit donc ici, de construire des entités enrichies disposant de propriétés que certaines entités n'ont pas dans le catalogue tout en identifiant des valeurs de propriétés agrégées, incomplètes voire erronées. Reprenons notre exemple du *génie des alpages*. Si l'on cherche des traductions de cette œuvre, on trouvera dans Wikidata que sa traduction norvégienne est *Ullkorn*. Cette information n'apparait que sur la source Wikidata, et n'existe pas non plus sur les catalogues des bibliothèques nationales. Comme le titre

traduit de cette bande-dessinée n'évolue pas, il est intéressant d'enrichir directement le catalogue avec un motif de dérivation comportant cette nouvelle information.

Le processus d'enrichissement dans un contexte d'héritage culturel peut se traiter de différentes manières. Les étapes d'alignement d'ontologies et de création de liens sont des pré-requis à l'enrichissement. L'ajout ou la fusion d'informations à partir d'une entité externe peut être perçu comme un processus de fusion de données (Dong *et al*., 2014 ; Brando *et al*., 2016), qui a été brièvement discuté en section 3.3.

Le catalogue suisse *linked.swissbib.ch* est, à notre connaissance, le seul à évoquer l'enrichissement (Bensmann *et al*., 2017). Pour rappel, le liage vers DBpedia et VIAF exploite quatre propriétés que sont le nom, le prénom, la date de naissance et celle de décès (voir section 4.3). Quand des liens sont découverts, les propriétés des entités liées sont ajoutées à l'entité équivalente du catalogue suisse lorsqu'elles sont absentes. Les auteurs ne décrivent pas comment l'alignement d'ontologies est réalisé pour les propriétés qui n'ont pas servi au liage (e.g., le métier de la personne). De plus, la fusion pour des propriétés équivalentes n'est pas évoquée.

L'un des enjeux est de sélectionner les propriétés ou valeurs à intégrer. DBpedia peut par exemple inclure de nombreuses propriétés peu pertinentes sur une personne dans le contexte bibliographique. Le projet Europeana intègre un processus d'augmentation automatique des données descriptives [76], mais dont les termes supplémentaires sont extraits de manière contrôlée du web des données (e.g., DBpedia, Geonames).

Afin de distinguer les données d'origine du catalogue de celles enrichies, il semble crucial de conserver une trace des informations à l'origine de l'entité enrichie, notamment dans le cas où ces informations évoluent (provenance). Par exemple, la solution LDIF transforme des données hétérogènes issues du web des données en une représentation cible propre et locale tout en gardant une trace de la provenance des données (Schultz *et al*., 2012).

Enfin, l'enrichissement peut concerner les réseaux sociaux, qui sont fortement utilisés par les bibliothèques pour disséminer des informations aux usagers. Un catalogue publié sous forme de données liées facilite la détection des ressources pertinentes pour promouvoir des événements sur des médias comme Twitter (Atefeh, Khreich, 2015). L'annonce de spectacles ou d'expositions peut donc s'accompagner d'informations plus détaillées provenant du catalogue.

5. Exploitation du catalogue sémantique

Une fois le catalogue migré selon un modèle sémantique, il est intéressant de considérer son exploitation et d'envisager de nouvelles fonctionnalités répondant aux besoins des utilisateurs, qu'ils soient bibliothécaires ou usagers de la bibliothèque. Dans cette section, nous nous intéressons aux problèmes d'indexation et de licence d'utili-

76. https://pro.europeana.eu/page/europeana-semantic-enrichment

sation, puis aux recherches sémantiques agrégeant dynamiquement différentes sources liées et ouvertes, et enfin à la visualisation et la navigation dans le nouveau catalogue.

5.1. Indexation

L'accès aux données LOD pose un problème d'indexation des données, entre autre par rapport aux moteurs de recherche comme Google. Ces derniers reconnaissent des vocabulaires comme `schema.org`, mais ne comprennent pas les vocabulaires plus spécifiques du monde bibliothécaire. Malgré l'existence d'un schéma spécifique aux livres (`bib.schema.org`), la communauté bibliographique, et notamment l'OCLC, est allée plus loin en proposant BiblioGraph[8], qui offre un compromis entre la vision générale des concepts du web de `schema.org` et la richesse des données bibliographiques. Cette extension devrait permettre à terme de mieux référencer les concepts des catalogues si elle est adoptée par les acteurs majeurs du web. Une autre solution pour éviter l'isolement des bibliothèques, consiste à aligner le vocabulaire vers d'autres plus génériques (e.g., Dublin Core, voire RDA). Comme le souligne Suominen *et al.*, un risque possible, si les bibliothèques ne s'accordent pas sur un modèle commun, est de voir un acteur majeur du web imposer un modèle, qui pourrait nuire à la qualité de la transformation (e.g., perte de données) ou être difficile à comprendre et utiliser (Suominen, Hyvönen, 2017).

5.2. Licence d'utilisation

Un autre problème concerne la gestion des droits sur les données publiées sous forme de LOD. Aujourd'hui, les bibliothèques sont rarement propriétaires des notices puisque celles-ci sont fournies et/ou vendues par des éditeurs ou des professionnels. Ces derniers devront donc publier leurs notices sous forme de données liées, mais un problème de propriété intellectuelle se pose, d'autant plus lorsque les bibliothécaires adaptent, corrigent ou enrichissent les notices. Comme les droits varient d'un pays à un autre, il est d'autant plus compliqué de collaborer pour publier ou intégrer des données ouvertes. La plupart des bibliothèques nationales (BL, BNF, DNB, BNE, LIBRIS, etc.) publient actuellement leur catalogue avec une licence « sans droit réservé » (Creative Commons 0), ce qui résout le problème de droits différents entre les pays (Chen, 2017).

5.3. Recherche sémantique

L'interrogation de catalogues sémantiques en lien avec des données ouvertes permet de considérer de nouvelles fonctionnalités apportant soit davantage d'informations en lien avec la recherche effectuée soit davantage de pertinence des résultats retournés grâce à une meilleure compréhension des requêtes utilisateurs.

Le principe des requêtes agrégatives est de pouvoir combiner dynamiquement les résultats issus de différentes sources pour fournir une vue d'ensemble à l'utilisateur

sur les différentes données en relation. Dans *data.bnf.fr* par exemple, les données initialement encodées en MARC, Dublin Core et EAD [77] et converties dans le même format du modèle RDF peuvent être interrogées *via* une seule requête. Ceci offre la possibilité au système de construire la page d'un auteur avec toutes les ressources liées à cet auteur. Autre exemple, l'interface de *linked.swissbib.ch* montre le résultat de l'agrégation de différentes informations pour un auteur ou une thématique, et offre des recommandations basées sur une proximité sémantique avec d'autres entités (Bensmann *et al*., 2017). Dans le processus d'identification d'une œuvre, la langue a une importance en particulier pour le traitement de requêtes portant sur des traductions. L'utilisation de données ouvertes offrent des perspectives intéressantes dans le cadre de recherche multilingues. Le portail *The European Library* [78] propose avec l'affichage du résultat d'une requête des sujets connexes au terme recherché que ce soit en français, en allemand ou en anglais. Le système exploite pour cela un processus d'alignement avec des fichiers d'autorités tels que RAMEAU [79], LCSH [16] et GND [80], pour lesquels l'alignement est issu du projet *Multilingual Access to Subjects* (Landry, 2004 ; Clavel-Merrin, 2004).

L'expression et la compréhension des requêtes des utilisateurs sont des problématiques à part entière, qui sont abordées dans notre contexte sous un angle de visualisation des motifs bibliographiques. Pour faciliter la formulation de requêtes, des interfaces de visualisation permettent de dessiner un graphe représentant cette requête (Zhu, Yan, 2016). Des fonctionnalités de correction et de désambiguïsation des requêtes complètent le système. L'article (Farrokhnia, Aalberg, 2016) s'intéresse quant à lui à la compréhension des besoins utilisateurs dans la recherche d'œuvres bibliographiques, en considérant notamment des motifs issus de cas d'utilisation.

5.4. Visualisation

De par la richesse de ses relations, la visualisation des informations culturelles est une problématique complexe (Windhager *et al*., 2016). Le système Ariadne [81] offre la possibilité de naviguer parmi 65 millions d'entités du catalogue OCLC représentant des articles scientifiques (Koopman *et al*., 2017). Un graphe est construit dynamiquement pour chaque requête et permet d'explorer le contexte associé à la requête. Contrairement au format MARC qui rendait difficile la sélection d'une œuvre, un catalogue sémantique identifie correctement les informations d'une œuvre. Mais le même défi se pose pour visualiser les éditions ou les formats de média. L'outil FRBRVis tente de résoudre ce problème en adaptant l'interface selon les différents concepts FRBR pour faciliter la navigation dans le catalogue (Aalberg *et al*., 2017). L'outil GlamMap permet de regrouper des œuvres selon leur emplacement géographique et de les vi-

77. https://www.loc.gov/ead/
78. http://www.theeuropeanlibrary.org/
79. http://rameau.bnf.fr/
80. http://www.dnb.de/DE/Standardisierung/GND/gnd_node.html
81. http://thoth.pica.nl/relate

sualiser sur une carte (Castermans *et al*., 2016). Là aussi, un modèle sémantique qui permet d'isoler les lieux (e.g., classe *Place*) est un atout pour ce type de visualisation dont l'objectif est de suivre l'évolution d'une ressource par exemple.

6. Conclusion et perspectives

L'arrivée du web sémantique est un bouleversement majeur pour la communauté bibliographique. Son système d'information inclut le catalogue des ressources disponibles, qui a peu évolué ces dernières décennies et dont les services associés (.e.g., recherche d'œuvres) sont fortement décriés (Gonzales, 2014). Pour rester une alternative crédible et pertinente aux moteurs de recherche, les bibliothèques doivent donc faire évoluer leur système d'information et la gestion de leur catalogue. La première étape consiste à changer le modèle de données, car des notices MARC qui décrivent l'ensemble des informations d'une œuvre (sans référence vers les entités impliquées) ne sont pas adaptées aux technologies du web sémantique. Plusieurs initiatives inspirées du modèle conceptuel FRBR (FRBRoo, RDA, BIBFRAME, etc.) font ou ont fait l'objet d'expérimentations, et les études soulignent déjà les avantages de ces modèles sémantiques en termes de navigation et d'enrichissement (Buchanan, 2006 ; Alemu *et al*., 2012). La seconde étape consiste à transformer les données MARC vers le modèle sélectionné (FRBRisation), puis de publier le catalogue résultant sous forme de données liées et ouvertes (Malmsten, 2008 ; Goddard, Byrne, 2010). De nombreux outils permettent l'interprétation des champs MARC *via* des correspondances, mais ils restent fragiles face aux pratiques locales de catalogage et aux champs de saisie libre. De plus, une difficulté majeure concerne l'interprétation des motifs bibliographiques comme les traductions ou adaptations (Aalberg *et al*., 2018). Après cette transformation, chaque entité bibliographique possède une URI, ce qui améliore la visibilité du catalogue (Prongué, Hügi, 2013). La disponibilité de sources de bonne qualité comme VIAF, DBpedia, ou Wikidata facilite la création de liens durables avec le catalogue tout en améliorant l'accessibilité aux ressources de la bibliothèque. Enfin, l'exploitation du catalogue implique une évolution des outils et services pour l'utilisateur des bibliothèques (e.g., visualisation améliorée des résultats, recherche sémantique, navigation entre les entités), mais également pour les bibliothécaires eux-mêmes (e.g., vérification des données suite à la transformation et au liage, assistance pour l'annotation des ressources ou enrichissement à partir de sources externes).

L'incubateur du W3C *Library Linked Data* publiait en 2011 les défis et recommandations [82] que les communautés bibliothèque et web sémantique doivent relever pour positionner les bibliothèques comme un acteur culturel de premier plan sur le web : modélisation des données, transformation des données originales, qualité des données, gestion des droits, collaboration avec les autres communautés. Sept années plus tard, ces pistes sont toujours d'actualité. Certes, de nombreuses institutions exposent désormais leur catalogue sous forme de LOD et de multiples expérimentations

82. http://www.w3.org/2005/Incubator/lld/XGR-lld/

proposent des solutions ou des améliorations pour résoudre certains verrous. Mais le nombre de modèles et de vocabulaires ne simplifient pas la modélisation ni la collaboration avec d'autres communautés, et leur spécificité exclue pour l'instant les données bibliographiques des résultats des moteurs de recherche. Plusieurs voix plaident pour une uniformisation des modèles, et pour l'adoption d'un vocabulaire simplifiant l'indexation sur le web. Pour l'interprétation des champs, l'exploitation de sources externes est peu répandue, car une majorité d'œuvres n'apparaissent pas sur le web des données. Pourtant, les notices les plus complexes à transformer sont en général les œuvres les plus populaires, c'est-à-dire celles qui comportent de nombreuses traductions et éditions. Le web des données, l'exposition des catalogues nationaux et celles des sources comme *WorldCat* ou *OpenLibrary* devraient faciliter l'interprétation des motifs bibliographiques pour ces œuvres populaires. Concernant la qualité de données, les bibliothécaires peuvent effectivement avoir des doutes par rapport à des informations qu'ils ne contrôlent pas, et qui sont parfois modifiables par le grand public (Goddard, Byrne, 2010). Le liage de données ou la réutilisation d'URI impliquent de faire confiance à la source qui les fournit. La maturité du web des données et des garanties de haut degré de qualité des données peuvent favoriser cette confiance. Enfin, les catalogues bibliographiques sont riches et globalement fiables, leurs données peuvent donc permettre de compléter, voire corriger, des informations présentes sur des sources externes. En particulier, certains concepts (expressions, manifestations) sont peu présents sur le LOD et les bibliothèques pourraient devenir fournisseur de données à ce niveau. L'une des perspectives les plus attendues porte sur une application qui valorise le catalogue sémantique et ses liens externes auprès des utilisateurs. Les bibliothèques manquent d'outils pour visualiser le catalogue FRBRisé et facilement corriger les erreurs pouvant résulter de la migration. L'apprentissage permettrait de reconnaitre puis de suggérer d'éventuelles erreurs. Il est envisageable que des communautés de bibliothécaires s'organisent pour partager et maintenir leurs catalogues, et que les usagers des bibliothèques s'investissent également *via* l'ajout collaboratif d'annotations. Enfin, l'utilisation du catalogue peut être analysée pour savoir comment sont consommées les données. Cette analyse peut porter sur le type de requêtes, les ressources demandées, les scénarios utilisateur, etc.

En conclusion, les bibliothèques évoluent lentement sur la publication de leur catalogue sous forme de données liées et ouvertes. Les expérimentations des grandes bibliothèques nationales ne portent que sur un sous-ensemble de leur catalogue MARC. Le manque de moyens et de personnel qualifié en web sémantique et en bibliographie est un frein à ce développement. Quand une application aura démontré le potentiel offert par les catalogues sémantiques, les bibliothécaires, chercheurs et industriels redoubleront sûrement d'efforts pour résoudre les derniers obstacles et permettre une migration et un liage semi-automatisé des catalogues de toutes les bibliothèques.

Remerciements

Ces travaux ont été en partie financés par l'Association Nationale de la Recherche et de la Technologie (ANRT, www.anrt.asso.fr), l'entreprise Progilone (www.progilone.com/), et un projet CNRS PICS (#PICS06945). Les auteurs remercient également les relecteur.e.s pour leurs commentaires et suggestions.

Bibliographie

Aalberg T. (2006). A Process and Tool for the Conversion of MARC Records to a Normalized FRBR Implementation. *ICADL*, vol. 4312, p. 283–292.

Aalberg T., Duchateau F., Takhirov N., Decourselle J., Lumineau N. (2018, 1). Benchmarking and evaluating the interpretation of bibliographic records. *International Journal on Digital Libraries (IJDL)*, p. 1-23. Consulté sur https://doi.org/10.1007/s00799-018-0233-2

Aalberg T., Merčun T., Žumer M. (2017). Interactive displays for the next generation of entity-centric bibliographic models. In *ICADL*, p. 199–211.

Aalberg T., Žumer M. (2013). The Value of MARC Data, or, Challenges of FRBRisation. *Journal of Documentation*, vol. 69, p. 851–872.

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

Atefeh F., Khreich W. (2015). A survey of techniques for event detection in Twitter. *Computational Intelligence*, vol. 31, n° 1, p. 132–164.

Baker T., Coyle K., Petiya S. (2014). Multi-entity models of resource description in the semantic web: A comparison of FRBR, RDA and BIBFRAME. *Library Hi Tech*, vol. 32, n° 4, p. 562–582.

Bensmann F., Zapilko B., Mayr P. (2017). Interlinking large-scale library data with authority records. *Frontiers in Digital Humanities*, vol. 4, p. 5.

Berners-Lee T., Hendler J., Lassila O. *et al*. (2001). The semantic web. *Scientific american*, vol. 284, n° 5, p. 28–37.

Billey A. M., L'Ecuyer-Coelho M.-C., Kovari J., Wacker M. (2018). *The Outcome of the Art-Frame Project, a Domain-Specific BIBFRAME Exploration*. Rapport technique. Columbia University Academic Commons. Consulté sur https://doi.org/10.7916/D8281M24

Bizer C., Heath T., Berners-Lee T. (2009). Linked data-the story so far. *Semantic services, interoperability and web applications: emerging concepts*, p. 205–227.

Bowen J. (2010). Moving Library Metadata Toward Linked Data: Opportunities Provided by the eXtensible Catalog. *Dublin Core and Metadata Applications*.

Brando C., Abadie N., Frontini F. (2016). Évaluation de la qualité des sources du web de données pour la résolution d'entités nommées. *ISI*, vol. 21, n° 5-6, p. 31–54.

Buchanan G. (2006). FRBR: Enriching and Integrating Digital Libraries. In *Joint Conference on Digital Libraries*, p. 260–269.

Bygstad B., Ghinea G., Klæboe G.-T. (2009). Organisational challenges of the Semantic Web in digital libraries: A Norwegian case study. *Online Information Review*, vol. 33, p. 973–985.

Candela G., Escobar P., Carrasco R. C., Marco-Such M. (2016). Migration of a library catalogue into RDA linked open data. *Semantic Web*, p. 1–11.

Castermans T., Speckmann B., Verbeek K., Westenberg M., Betti A., Berg H. van den. (2016). GlamMap: geovisualization for e-humanities. In *Visualization for the digital humanities.*

Chen. (2006). MetaLib, WebFeat, and Google: The strengths and weaknesses of federated search engines compared with Google. *Online Information Review*, vol. 30, p. 413–427.

Chen. (2017). A Review of Practices for Transforming Library Legacy Records into Linked Open Data. In *Research conference on metadata and semantics research*, p. 123–133.

Choffé P., Leresche F. (2016). DOREMUS: Connecting Sources, Enriching Catalogues and User Experience. *IFLA World Library and Information Congress*. Consulté sur http://library.ifla.org/1322/

Christen P. (2012a). *Data matching: concepts and techniques for record linkage, entity resolution, and duplicate detection*. Springer Science & Business Media.

Christen P. (2012b). A survey of indexing techniques for scalable record linkage and deduplication. *IEEE transactions on knowledge and data engineering*, vol. 24, n° 9, p. 1537–1555.

Clavel-Merrin G. (2004). MACS (Multilingual access to subjects): a virtual authority file across languages. *Cataloging & Classification Quarterly*, vol. 39, n° 1-2, p. 323–330.

Cole T. W., Han M.-J., Weathers W. F., Joyner E. (2013). Library MARC records into Linked open data: challenges and opportunities. *Journal of Library Metadata*, vol. 13, p. 163–196.

Committee S., Group I. S. (1998). *Functional Requirements for Bibliographic Records: final report* (vol. 19). K. G. Saur.

Consoli S., Recupero D. R. (2015). Using FRED for named entity resolution, linking and typing for knowledge base population. In *Semantic web evaluation challenge*, p. 40–50.

Coyle K. (1992). *Rules for Merging MELVYL Records. Technical Report No. 6. Revised*. ERIC.

Coyle K. (2014). FRBR, Twenty Years On. *Cataloging & Classification Quarterly*, p. 1–21.

Coyle K. (2016). *FRBR before and after*. ALA.

Cronin C. (2011). From testing to implementation: Managing full-scale RDA adoption at the University of Chicago. *Cataloging & Classification Quarterly*, vol. 49, n° 7-8, p. 626–646.

Cruz J. M. B., Testal C. G. (2013). Application of LOD to Enrich the Collection of Digitized Medieval Manuscripts at the University of Valencia. *SWIB Conference*. Consulté sur http://swib.org/swib13/

Decourselle J., Duchateau F., Lumineau N. (2015). A Survey of FRBRization Techniques. In *Theory and Practice of Digital Libraries (TPDL)*, p. 185-196.

Deliot C. (2014). Publishing the British national bibliography as linked open data. *Catalogue & Index*, vol. 174, p. 13–18.

Doerr M., Gradmann S., Hennicke S., Isaac A., Meghini C., Sompel H. van de. (2010). The Europeana data model (EDM). In *Ifla world library and information congress*, p. 10–15.

Dong X. L., Gabrilovich E., Heitz G., Horn W., Murphy K., Sun S. *et al*. (2014). From data fusion to knowledge fusion. *Proc. VLDB Endow.*, vol. 7, n° 10, p. 881–892.

Duchateau F., Takhirov N., Aalberg T. (2011). FRBRPedia: a Tool for FRBRizing Web Products and Linking FRBR Entities to DBpedia. In *Joint Conference on Digital Libraries*, p. 455-456.

Euzenat J., Shvaiko P. (2013). *Ontology Matching*. Springer Science & Business Media.

Farrokhnia M., Aalberg T. (2016). Finding user need patterns in the world of complex semantic cultural heritage data. In *Metadata and semantics research*, p. 187–192.

Fellegi I. P., Sunter A. B. (1969). A theory for record linkage. *Journal of the American Statistical Association*, vol. 64, n° 328, p. 1183–1210.

Freire N., Borbinha J., Calado P. (2007). Identification of FRBR works within bibliographic databases: An experiment with UNIMARC and duplicate detection techniques. In *International conference on asian digital libraries*, p. 267–276.

Furrie B. (2000). Understanding MARC bibliographic machine readable cataloging.

Gandon F., Corby O., Faron-Zucker C. (2012). *Le web sémantique: Comment lier les données et les schémas sur le web?* Dunod.

Gibson I., Goddard L., Gordon S. (2009). One box to search them all: Implementing federated search at an academic library. *Library Hi Tech*, vol. 27, n° 1, p. 118–133.

Goddard L., Byrne G. (2010). The strongest link: Libraries and linked data. *D-Lib magazine*, vol. 16, n° 11/12.

Gonzales B. M. (2014). Linking libraries to the web: linked data and the future of the bibliographic record. *Information Technology and Libraries (Online)*, vol. 33, n° 4, p. 10.

Hallo M., Luján-Mora S., Maté A., Trujillo J. (2016). Current state of Linked Data in digital libraries. *Journal of Information Science*, vol. 42, n° 2, p. 117–127.

Hallo M., Luján-Mora S., Trujillo J. (2014). Transforming library catalogs into Linked Data. In *Iceri proceedings*, p. 1845-1853. IATED.

Hammerton J. A., Granitzer M., Harvey D., Hristakeva M., Jack K. (2012). On generating large-scale ground truth datasets for the deduplication of bibliographic records. In *International conference on web intelligence, mining and semantics*, p. 18.

Han M.-J. K., Cole T. W., Sarol M. J., Lampron P., Wade M., Stacker T. *et al.* (2016). Linked Open Data in Practice: Emblematica Online. *SWIB Conference*. Consulté sur http://swib.org/swib16/

Hannemann J., Kett J. (2010). Linked data for libraries. In *IFLA*.

Harper C. A., Tillett B. B. (2007). Library of Congress controlled vocabularies and their application to the Semantic Web. *Cataloging & classification quarterly*, vol. 43, n° 3-4, p. 47–68.

Haslhofer B., Isaac A. (2011). data.europeana.eu: The europeana linked open data pilot. In *Dublin core and metadata applications*, p. 94–104.

He W., Mihara T., Nagamori M., Sugimoto S. (2013). Identification of Works of Manga Using LOD Resources: An Experimental FRBRization of Bibliographic Data of Comic Books. In *Joint conference on digital libraries*, p. 253–256.

Hickey T. B., O'Neill E. T. (2005). FRBRizing OCLC's WorldCat [FRBRization (OCLC)]. *Cataloging & Classification Quarterly*, vol. 39, p. 239–251.

Hladka J., Mynarz J., Sklenak V. (2012). Experience with transformation of bibliographic data into linked data. *Journal of Systems Integration*, vol. 3, n° 1, p. 54.

Jentzsch A., Isele R., Bizer C. (2010). SILK - generating rdf links while publishing or consuming linked data. In *ISWC*, p. 53–56.

Kaenel I. de, Iriarte P. (2007). Les catalogues des bibliothèques: du web invisible au web social. *RESSI: Revue électronique suisse de science de l'information*, n° 5.

Kempf A., Neubert J. (2016, 04). The Role of Thesauri in an Open Web: A Case Study of the STW Thesaurus for Economics. , vol. 43, p. 160-173.

Koopman R., Wang S., Scharnhorst A. (2017). Contextualization of topics: browsing through the universe of bibliographic information. *Scientometrics*, vol. 111, n° 2, p. 1119–1139.

Kovari J., Folsom S., Younes R. (2017). Towards a BIBFRAME implementation: the bibliotek-o framework. In *Dublin core and metadata applications*, p. 52–61.

Krafft D. B. (2015). Linked data for libraries: a project update. In *ISWC*.

Kroeger A. (2013). The road to BIBFRAME: the evolution of the idea of bibliographic transition into a post-MARC Future. *Cataloging & classification quarterly*, vol. 51, n° 8, p. 873–890.

Lampert C. K., Southwick S. B. (2013). Leading to linking: Introducing linked data to academic library digital collections. *Journal of Library Metadata*, vol. 13, n° 2-3, p. 230–253.

Landry P. (2004). Multilingual subject access: The linking approach of MACS. *Cataloging & Classification Quarterly*, vol. 37, n° 3-4, p. 177–191.

Latif A., Borst T., Tochtermann K. (2014). Exposing data from an open access repository for economics as linked data. *D-Lib Magazine*, vol. 20, n° 9/10.

Luzzi C. (2014). ManUScript Italian poEtry in muSic (1500-1700) interoperable model: towards an application of FRBRoo, Linked Open Data and Semantic Web technology. In *Workshop on digital libraries for musicology*, p. 1–3.

Malmsten M. (2008). Making a library catalogue part of the semantic web. In *Dublin core and metadata applications.*

Marketakis Y., Minadakis N., Kondylakis H., Konsolaki K., Samaritakis G., Theodoridou M. *et al*. (2017). X3ML mapping framework for information integration in cultural heritage and beyond. *International Journal on Digital Libraries*, vol. 18, n° 4, p. 301–319.

Mayr P., Petras V. (2008). Cross-concordances: terminology mapping and its effectiveness for information retrieval.

Mazurek C., Sielski K., Walkowska J., Werla M. (2012). From MARC21 and Dublin Core, through CIDOC CRM: First Tenuous Steps towards Representing Library Data in FRBRoo. *CIDOC 2012*.

McGee M., Durante K., Weimer K. H. (2017). Toward a Linked Data Model for Describing Cartographic Resources. *Journal of Map & Geography Libraries*, vol. 13, n° 1, p. 133-144.

Newcombe H. B., Kennedy J. M., Axford S., James A. P. (1959). Automatic linkage of vital records. *Science*, p. 954–959.

Otero-Cerdeira L., Rodríguez-Martínez F. J., Gómez-Rodríguez A. (2015). Ontology matching: A literature review. *Expert Systems with Applications*, vol. 42, n° 2, p. 949–971.

Phipps J., Dunsire G., Hillmann D. (2015). Building a Platform to Manage RDA Vocabularies and Data for an International Linked Data World. *Journal of Library Metadata*, vol. 15, p. 252-264.

Prongué N., Hügi J. (2013). Les applications basées sur les lod en bibliothèque. *Arbido*, n° 3, p. 15–16.

Raimond Y., Sutton C., Sandler M. B. (2008). Automatic Interlinking of Music Datasets on the Semantic Web. *LDOW*, vol. 369.

Riva P. (2004). Mapping MARC 21 Linking Entry Fields to FRBR and Tillett's Taxonomy of Bibliographic Relationships. *Library resources & technical services*, vol. 48, p. 130–143.

Riva P., Le Boeuf P., Žumer M. (2016). *FRBR-Library Reference Model*. Rapport technique. IFLA FRBR Review Group.

Sadeh T. (2007). Time for a change: new approaches for a new generation of library users. *New Library World*, vol. 108, n° 7/8, p. 307–316.

Schultz A., Matteini A., Isele R., Mendes P. N., Bizer C., Becker C. (2012). LDIF - a framework for large-scale linked data integration. In *World wide web conference.*

Shen W., Wang J., Han J. (2015). Entity linking with a knowledge base: Issues, techniques, and solutions. *Transactions on Knowledge and Data Engineering*, vol. 27, n° 2, p. 443–460.

Simon A., Wenz R., Michel V., Di Mascio A. (2013). Publishing bibliographic records on the Web of data: Opportunities for the BnF (French National Library). In *ESWC*, p. 563–577.

Sitas A., Kapidakis S. (2008). Duplicate detection algorithms of bibliographic descriptions. *Library hi tech*, vol. 26, n° 2, p. 287–301.

Suominen O. (2017). Finnish National Bibliography Fennica as Linked Data. *SWIB Conference*. Consulté sur http://swib.org/swib17/

Suominen O., Hyvönen N. (2017). From MARC silos to Linked Data silos? *o-bib*, vol. 4, n° 2. Consulté sur https://www.o-bib.de/article/view/2017H2S1-13

Szekely P., Knoblock C. A., Yang F., Zhu X., Fink E. E., Allen R. *et al.* (2013). Connecting the smithsonian american art museum to the linked data cloud. In *ESWC*, p. 593–607.

Takhirov N., Aalberg T., Duchateau F., Žumer M. (2012). FRBR-ML: A FRBR-based framework for semantic interoperability. *Semantic Web Journal*, vol. 3, p. 23–43.

Takhirov N., Duchateau F., Aalberg T. (2011). Linking FRBR Entities to LOD through Semantic Matching. In *Theory and Practice of Digital Libraries (TPDL)*, p. 284-295.

Teets M., Goldner M. (2013). Libraries' Role in Curating and Exposing Big Data. *Future Internet*, vol. 5, n° 3, p. 429–438. Consulté sur http://www.mdpi.com/1999-5903/5/3/429

Tennant R. (2002). MARC must die. *Library Journal*, vol. 127, n° 17, p. 26–27.

Vila-Suero D., Villazón-Terrazas B., Gómez-Pérez A. (2013). datos.bne.es: A library linked dataset. *Semantic Web*, vol. 4, n° 3, p. 307–313.

Volz J., Bizer C., Gaedke M., Kobilarov G. (2009). Discovering and maintaining links on the web of data. In *International semantic web conference*, p. 650–665.

Wang Y., Dawes T. A. (2012). The next generation integrated library system: a promise fulfilled. *Information Technology and Libraries (Online)*, vol. 31, n° 3, p. 76.

Westrum A.-L., Rekkavik A., Tallerås K. (2012). Improving the presentation of library data using FRBR and Linked data. *Code4Lib Journal*, vol. 16, n° 0.

Windhager F., Federico P., Mayr E., Schreder G., Smuc M. (2016). A review of information visualization approaches and interfaces to digital cultural heritage collections. In *Proceedings of Forum Media Technology*, p. 23–24.

Zapounidou S., Sfakakis M., Papatheodorou C. (2014). Library data integration: towards BIBFRAME mapping to EDM. In *Metadata and semantics research*, p. 262–273.

Zapounidou S., Sfakakis M., Papatheodorou C. (2017). Preserving Bibliographic Relationships in Mappings from FRBR to BIBFRAME 2.0. In *TPDL*, p. 15–26.

Zhu Y., Yan E. (2016). Searching bibliographic data using graphs: A visual graph query interface. *Journal of Informetrics*, vol. 10, n° 4, p. 1092 - 1107.