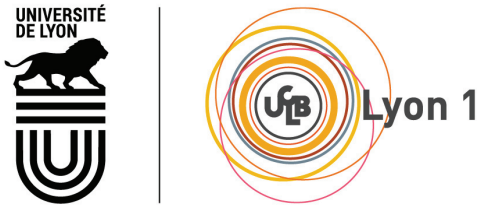
Migration et enrichissement sémantique d'entités culturelles

Joffrey Decourselle



N° d'ordre NNT : 2018LYSE1183

THÈSE DE DOCTORAT DE L'UNIVERSITÉ DE LYON
opérée au sein de
l'Université Claude Bernard Lyon 1

École Doctorale : **ED512 InfoMaths**

Discipline : **Informatique**
Spécialité : **Science des données**

soutenue le 28/09/2018
par :
Joffrey Decourselle

Migration et Enrichissement Sémantique d'Entités Culturelles

Composition du jury :

BOUZEGHOUB Amel, Professeure, Télécom SudParis	Rapporteure
HAMEURLAIN Abdelkader, Professeur, Université Toulouse 3	Rapporteur(e)
ZEITOUNI Karine, Professeure, Université de Versailles Saint-Quentin	Examinatrice
HACID Mohand-Saïd, Professeur, Université Lyon 1	Directeur de thèse
LUMINEAU Nicolas, Maître de Conférences, Université Lyon 1	Co-directeur de thèse

UNIVERSITE CLAUDE BERNARD - LYON 1

Président de l'Université	**M. le Professeur Frédéric FLEURY**
Président du Conseil Académique	M. le Professeur Hamda BEN HADID
Vice-président du Conseil d'Administration	M. le Professeur Didier REVEL
Vice-président du Conseil Formation et Vie Universitaire	M. le Professeur Philippe CHEVALIER
Vice-président de la Commission Recherche	M. Fabrice VALLÉE
Directrice Générale des Services	Mme Dominique MARCHAND

COMPOSANTES SANTE

Faculté de Médecine Lyon Est – Claude Bernard	Directeur : M. le Professeur G.RODE
Faculté de Médecine et de Maïeutique Lyon Sud – Charles Mérieux	Directeur : Mme la Professeure C. BURILLON
Faculté d'Odontologie	Directeur : M. le Professeur D. BOURGEOIS
Institut des Sciences Pharmaceutiques et Biologiques	Directeur : Mme la Professeure C. VINCIGUERRA
Institut des Sciences et Techniques de la Réadaptation	Directeur : M. X. PERROT
Département de formation et Centre de Recherche en Biologie Humaine	Directeur : Mme la Professeure A-M. SCHOTT

COMPOSANTES ET DEPARTEMENTS DE SCIENCES ET TECHNOLOGIE

Faculté des Sciences et Technologies	Directeur : M. F. DE MARCHI
Département Biologie	Directeur : M. le Professeur F. THEVENARD
Département Chimie Biochimie	Directeur : Mme C. FELIX
Département GEP	Directeur : M. Hassan HAMMOURI
Département Informatique	Directeur : M. le Professeur S. AKKOUCHE
Département Mathématiques	Directeur : M. le Professeur G. TOMANOV
Département Mécanique	Directeur : M. le Professeur H. BEN HADID
Département Physique	Directeur : M. le Professeur J-C PLENET
UFR Sciences et Techniques des Activités Physiques et Sportives	Directeur : M. Y.VANPOULLE
Observatoire des Sciences de l'Univers de Lyon	Directeur : M. B. GUIDERDONI
Polytech Lyon	Directeur : M. le Professeur E.PERRIN
Ecole Supérieure de Chimie Physique Electronique	Directeur : M. G. PIGNAULT
Institut Universitaire de Technologie de Lyon 1	Directeur : M. le Professeur C. VITON
Ecole Supérieure du Professorat et de l'Education	Directeur : M. le Professeur A. MOUGNIOTTE
Institut de Science Financière et d'Assurances	Directeur : M. N. LEBOISNE

Remerciements

L'aboutissement de cette thèse n'aurait pu avoir lieu sans la bienveillance de mes encadrants Nicolas Lumineau, Fabien Duchateau et Mohand-Saïd Hacid et la confiance de Michel Vivier, directeur général de la société Progilone, que je tiens à remercier très chaleureusement. Tous auront su créer un environnement idéal pour la réalisation de cette thèse CIFRE dans un domaine particulièrement balancé entre enjeux scientifiques et enjeux techniques. Je souhaite également remercier les rapporteurs Amel Bouzeghoub, Abdelkader Hameurlain et l'examinatrice Karine Zeitouni d'avoir accepté d'évaluer ce travail de thèse. J'ai une pensée toute particulière pour Trond Aalberg, professeur à l'Université de Sciences et Technologies (NTNU) de Norvège qui, dans le cadre du projet PICS Franco-Norvégien DIRICKS, a très fortement contribué à la bonne orientation des travaux réalisés. Cette collaboration scientifique avec la Norvège et le partenariat industriel avec l'entreprise Progilone ont facilité la concrétisation d'idées ambitieuses et pragmatiques dans le domaine documentaire. Je terminerai en remerciant tous ceux qui ont apporté leur pierre dans ce projet et qui m'ont surtout aidé à avancer, notamment ma famille, mes proches et mes amis Rémy, Roland, Mehdi, Cédric, etc. Cette thèse est dédiée à mes rayons de soleil : Zoé, Christelle, Yolaine et Fanny.

Résumé

Migration et Enrichissement Sémantique d'Entités Culturelles

Les technologies du Web Sémantique offrent de nouvelles possibilités pour améliorer les services dédiés à la diffusion des connaissances culturelles et intellectuelles. Cependant, de nombreuses institutions, en charge de fonds documentaires, doivent gérer leurs catalogues selon des normes héritées des années 60. Le paradigme historique des *notices papier*, où chaque notice décrit un document possédé par une institution, est encore majoritairement utilisé par la communauté documentaire. Les notices, aujourd'hui numériques, sont toujours conçues dans le but d'être lues et comprises par des documentalistes. Ainsi, elles reposent sur des modèles spécifiques à ce métier qui ne permettent pas une réutilisation idéale des connaissances qu'elles contiennent. Dans ce contexte, les nombreux catalogues contenant ces notices demeurent isolés du mouvement actuel prônant l'interopérabilité et l'exploration sémantique des sources de données. C'est pourquoi, différentes approches ont été étudiées depuis plus de 20 ans pour rendre possible l'adoption des technologies sémantiques dans l'univers bibliographique. De nouveaux modèles comme FRBR/LRM, ont été développés par la communauté pour permettre l'organisation des informations documentaires selon le paradigme d'entités et d'associations et de nouveaux vocabulaires sémantiques comme RDA facilitent l'interopérabilité des bases de données bibliographiques avec le web de données. Toutefois, un enjeu majeur pour les institutions documentaires consiste à transformer l'ensemble des données bibliographiques existantes, issues des anciens paradigmes, vers ces nouveaux modèles et vocabulaires sémantiques.

L'interprétation et la migration des anciennes notices bibliographiques vers des bases de connaissances sémantiques implique de relever des enjeux scientifiques importants. Un premier challenge consiste à adapter les modèles du domaine documentaire aux formalismes et principes du web sémantique. En effet, le patrimoine bibliographique est riche de multiples relations entre les documents permettant la description de familles bibliographiques complexes dans un catalogue documentaire. Cette richesse bibliographique, peu considérée par la communauté du web sémantique, doit être modélisée et intégrée dans les expérimentations de cette communauté avec des technologies adaptées. En ce sens, on observe un manque de jeux de données et de métriques qui intègrent ces relations riches et particulières au domaine documentaire. Un autre challenge est celui de l'interprétation des connaissances bibliographiques, issues des notices, avant de pouvoir les intégrer dans de nouvelles bases de données. La complexité de cette tâche d'interprétation peut varier selon les pratiques de catalogage des institutions documentaires et les modèles utilisées par ces dernières. Considérant que certains catalogues contiennent de très nombreuses notices, leur transformation implique le développement d'outils automatiques qui considèrent d'un côté les relations riches et spécifiques du domaine documentaires et d'un autre côté qui s'adaptent aux pratiques spécifiques des institutions. Pour résumer, l'adoption des technologies du web sémantique dans la communauté documentaire est partagée entre un processus complexe et long de normalisation et d'évaluation des connaissances bibliographiques pour respecter la qualité et richesse des données et un besoin immédiat d'outils permettant la transformation des catalogues existants. Ce double-enjeu implique de nombreux efforts aux institutions documentaires pour entamer leur conversion vers le web sémantique. Bien que plusieurs agences nationales aient initié des démarches de transformation de leurs catalogues, de nombreuses institutions publiques

comme privées, possédant des ressources plus spécialisées ou exclues des catalogues nationaux, manquent de compétences techniques et d'outils adaptées pour réussir une transformation satisfaisant les enjeux de qualité du domaine. Face à cette problématique, l'enjeu principal de cette thèse consiste à apporter des solutions innovantes pour la migration et l'enrichissement des catalogues bibliographiques pour former de nouvelles bases de connaissances.

La première contribution de cette thèse concerne l'évaluation de la qualité des bases de connaissances bibliographiques. En effet, l'évolution des données bibliographiques vers de nouvelles bases de connaissances sémantiques doit respecter les exigences de la communauté en termes de qualité et de réutilisabilité des données. Cependant, les écarts importants entre les anciens modèles de notices et les ontologies du web sémantique soulèvent des divergences dans la communauté concernant la modélisation des connaissances et la standardisation des nouveaux modèles bibliographiques. A cela s'ajoute le manque de métriques et d'expérimentations rendant difficile l'évaluation des systèmes et outils informatiques utilisés dans ce domaine. C'est pourquoi nous proposons un benchmark original qui est dédié à l'interprétation et à la transformation des catalogues bibliographiques ainsi qu'à l'évaluation des nouvelles bases de connaissances sémantiques émergentes. Ce benchmark est composé d'une part de métriques qui permettent d'anticiper les efforts de transformation des catalogues et de faciliter la création de nouvelles solutions informatiques dédiées à ce processus. D'autre part, les jeux de données que nous proposons intègrent un ensemble exhaustif de tests sur les spécificités des notices bibliographiques.

La deuxième contribution de cette thèse est une méthodologie pour l'extraction automatique des connaissances avancées d'un catalogue bibliographique. Notre objectif est de faciliter la création, par les documentalistes, de modèles de règles qui permettent l'interprétation et la transformation des notices en considérant les particularités du domaine bibliographique. Notre méthode considère notamment l'extraction des motifs de connaissances avancés (*ex.*, traductions, agrégations, illustrations) en bénéficiant de mécanismes apportés par les graphes d'entités et d'associations bibliographiques. Notre approche repose sur l'hypothèse que la transformation de connaissances complexes peut être simplifiée par la gestion de règles de migration à un niveau d'abstraction élevé. Cela signifie que notre système propose une gestion des règles au niveau des motifs de connaissances bibliographiques de la communauté et pas au niveau des entités des modèles existants comme c'est le cas dans des outils de migration plus courants. L'originalité de notre approche est de faciliter l'écriture de règles de migration et d'enrichissement des notices tout en améliorant la qualité globale du processus grâce à une meilleurs lisibilité des règles et par la prise en compte des relations bibliographiques avancées du domaine documentaire.

La troisième contribution s'inscrit dans la continuité des deux contributions précédentes. Nous proposons l'implémentation d'un système d'intégration de données bibliographiques qui repose nativement sur les métriques de qualité du domaine, pour mieux interpréter les catalogues à migrer, ainsi que notre méthodologie de modélisation des règles de transformation pour faciliter le processus aux experts documentaires. Notre solution permet la modélisation de règles, à un niveau d'abstraction élevé, pour réaliser la migration des notices et leur enrichissement à partir de sources de données externes. L'objectif principal de notre système est de concilier le besoin de flexibilité dans la conception des futurs modèles de bases de connaissances des institutions avec la nécessité d'avoir un outil opérationnel pour transformer des catalogues de manière automatique. Nous présentons les caractéristiques de notre solution ainsi que des résultats préliminaires d'expérimentations dans des contextes réels et industriels. L'objectif principal de ces travaux est de faciliter la réalisation de nouvelles solutions informatiques, dans la communauté documentaire, qui soient en phase avec les perspectives et objectifs scientifiques de cette communauté et aussi avec les enjeux concrets des professionnels du domaine.

Abstract

Migration and Semantic Enrichment of Cultural Entities

Semantic web technologies provide new solutions to enhance the services dedicated to the reuse of cultural heritage data on the web. However, many cultural institutions, in charge of documentary catalogs, still manage their data using former norms and systems. The old *card catalogs* which have been dedicated to the storage of bibliographic metadata during decades is still largely used by digital libraries. Digital bibliographic records are still made to be read and understood by documentary experts. Thus, these records still rely on former and specific models which prevent any reuse of the knowledge they describe. In such a context, many bibliographic catalogs, bases on these records remain isolated from the semantic web movement towards interoperability between knowledge bases. That is why various approaches have been studied in the bibliographic community to ease the adoption of semantic technologies in documentary institutions. New models like FRBR/LRM have been developed in the community to model the bibliographic knowledge using the entity and relationship paradigm. Moreover, vocabularies like RDA ease the description of bibliographic metadata in the semantic web. However, a major challenge for cultural institutions consists in transforming the large amount of existing bibliographic records towards these new semantic models and vocabularies.

The interpretation and migration of former bibliographic records, towards new semantic knowledge bases, bring several scientific challenges. A first challenge is the adaptation of existing bibliographic models with the semantic web formalisms and principles. Indeed, the bibliographic knowledge from an institution may describe rich relationships between documents leading to complex bibliographic families. Such rich information, not much considered in the semantic web community, must be modeled and integrated to the experimentations and tools from this community. Yet, we observe a lack of datasets and metrics which consider these rich relationships from the documentary domain. Another challenge is the interpretation of bibliographic knowledge in order to migrate such data into new semantic knowledge bases. The complexity of such task may vary according to the cataloging practices of institutions and the models the latter are using. Taking into account that bibliographic catalogs may contain many records, their transformation implies to develop new automated tools which consider both the rich relationships between documents and the specific cataloging practices from institutions. All in all, the adoption of semantic web technologies is at the frontier between the efforts of normalization and evaluation of the rich bibliographic knowledge using semantic technologies and the immediate need to get tools that allow the automated transformation of existing catalogs. This double challenge implies many efforts for cultural institutions to start their conversion towards semantic web principles. Although some national institutions have started the transformation of their catalogs, many institutions which own more specific resources, lack of technical skills and suitable tools to make a success of the transformation of their catalogs in respect to the quality challenges of the domain. Hence, the main chalenge we consider in this thesis is to bring innovative solutions for the migration and the semantic enrichment of bibliographic catalogs to create new semantic knowledge bases.

The first contribution of this thesis concerns the qualitative evaluation of bibliographic knowledge bases. Indeed, the data integration of bibliographic knowledge must respect the requirements of the community in terms of quality and reusability of the data. Yet, the differences between old models and new ontologies lead to various way to model some information making it more difficult to evaluate the quality of new knowledge bases. Moreover, the lack of metrics and of experimentations make it difficult the evaluation of IT tools in this domain. Hence, we provide an original benchmark dedicated both to the interpretation and transformation of bibliographic catalog and to the evaluation of the new semantic knowledge bases. Our benchmark is composed of metrics which allow to anticipate the efforts related to the transformation of catalogs, and of datasets which include a set of tests related to the specificities of bibliographic records.

The second contribution is a methodology for the automated extraction of the advanced knowledge from a bibliographic catalog. Our objective is to ease the creation, by documentary experts, of the rules required to interpret and transform the records with respect to the specifications of the bibliographic domain. Our method considers the extraction of advanced bibliographic patterns (e.g., transformations, aggregations, illustrations) by taking advantage of the mechanisms brought by graphs of entities and relationships from the semantic web. Our approach also relies on the hypothesis that the transformation of complex knowledge should be managed with high level patterns instead of entity-centric rules. This means that our system is based on bibliographic knowledge patterns which ease the task of documentary expert to model the knowledge for their futures databases. Thus, our proposal must ease the writing task of the rules to transform the metadata while improving the global quality of such process thanks to a better readability of the rules and to the consideration of the advanced bibliographic relationships from this domain.

Our third contribution follows the two previous contributions. We describe an implementation of a data integration system for bibliographic resources which both handle the qualitative metrics of the domain and our methodology to model the transformation rules. Our solution, based both on high leveled knowledge patterns and data integration solutions, ease the modeling task of rules to both migrate the records to a semantic model and to enrich semantically the data using external sources of data. Hence, our system aims to conciliate both the need from expert for more flexibility to build new knowledge models with another need by institutions to get practical tools which automatically transform existing catalogs. We present the characteristics of our solution and we provide preliminary results of experiments made in real and industrial contexts. Our works should ease the development of new solutions in the documentary community which take into account the scientific requirements of the domain and the challenges raised by the institutions.

Table des matières

Table des figures

Table des définitions

Chapitre 1

Introduction

"Le lieu d'emmagasinement et de classement devient aussi un lieu de distribution, [...]. De là on fait apparaître sur l'écran la page à lire pour connaître la réponse aux questions posées [...]. Utopie aujourd'hui parce qu'elle n'existe encore nulle part, mais elle pourrait bien devenir la réalité de demain pourvu que se perfectionnent encore nos méthodes et notre instrumentation."

Paul Otlet, *Traité de Documentation*, 1934

Sommaire

Les ambitions technologiques des sociétés modernes reposent de plus en plus sur une utilisation des données informatiques qui proviennent de multiples sources comme des réseaux sociaux, des images médicales, des capteurs connectés ou d'autres statistiques scientifiques. Une des problématiques majeures pour l'exploitation de ces données est leur *modélisation* afin de comprendre « leur nature et leur variabilité » (S. Mallat). Ce besoin de modèles intellectuels pour comprendre nos données est par ailleurs la source de toute la complexité que soulève le domaine de l'intelligence artificielle qui est discuté dans la société. Dans ce mouvement, les technologies du Web Sémantique et notamment le *web de données* sont reconnus pour apporter des solutions efficaces pour une meilleure représentation des données dans différents domaines [53].

Cette voie est d'ailleurs le point de départ d'un grand renouveau dans la communauté des humanités numériques [124]. Cette dernière, discutant des problématiques de gestion et de valorisation du patrimoine culturel de l'humanité, avec les technologies numériques, étudie notamment les enjeux de la modélisation informatique des ressources culturelles à disposition des étudiants, chercheurs ou du grand publique. Dans ce contexte, la thèse que nous présentons dans ce manuscrit décrit une méthodologie permettant de faciliter l'évolution des données bibliographiques issues de notre héritage culturel vers des systèmes d'information utilisant les principes et technologies du web de données. Dans le reste de ce premier chapitre, nous décrivons successivement le contexte de notre travail de recherche, les différentes problématiques que nous avons étudiées et, enfin, nous livrons une vision synthétique du travail réalisé ainsi que son contexte d'application.

1.1 Contexte de la thèse

Le développement de l'humanité repose sur la capacité des Hommes à transmettre les savoirs et connaissances aux générations futures. A travers les siècles, cette transmission a été bouleversée par de grands changements technologiques comme l'écriture ou l'imprimerie, transformant au passage les besoins en matière de sauvegarde et de diffusion du patrimoine intellectuel de l'humanité. Jusque dans les années 1940, la préservation et consultation des ressources bibliographiques était universellement assurée par des notices papiers (*cf.*, travaux de Paul Otlet [98]) contenant les informations sur chaque document comme illustré par la figure 1.1. Dans la communauté documentaire, l'arrivée du numérique et l'explosion des publications littéraires et électroniques a également imposé une évolution des méthodes et des outils dédiés à la préservation et gestion du patrimoine bibliographique. Dans ce contexte, les anciennes et universelles notices papier, décrivant les ressources documentaires (*ex.*, livre, journal) ont été progressivement transformées en notices numériques contenant les métadonnées bibliographiques de ces mêmes ressources. Les règles de catalogage ont également subi cette transformation vers le numérique. Enfin, de multiples systèmes d'information ont été développés pour répondre aux besoins de stockage et de préservation du patrimoine culturel avec des technologies numériques.

FIGURE 1.1 – Catalogue de notices bibliographiques entre 1900 et 1920. Photographie de la bibliothèques du congrès, D.C. 20540 USA, Reproduction LC-USZ62-118630

Cependant, le paradigme des "notices bibliographiques" a été régulièrement contesté, notamment pour ses limites concernant la réutilisation des métadonnées [110, 131, 30, 85]. En effet, le format de notices le plus répandu, nommé MARC, utilise un vocabulaire et une codification très spécifique au métier bibliographique [85], rendant difficile la réutilisation des données par d'autres acteurs. De plus, MARC est principalement dédié à la représentation de publications littéraires (ex., livres) comme on pourrait les imaginer sur un étagère. Ce paradigme reste malheureusement loin de la réalité numérique grandissante (ex., ebooks, web séries, podcasts) et des possibilités du Web Sémantique pour décrire des relations riches entre des documents (ex., adaptations, traductions, illustrations) [70]. Lee et Jacob résument ces problèmes en écrivant « *the reality is that the [MARC] format [...] has been primarily intended for the exchange and*

display of records and the data was, to a large degree, structured for human interpretation and not for automated processing and retrieval as it is required today » [80]. Les conséquences de ces limitations s'observent par l'éloignement des services numériques de certaines bibliothèques des nouvelles technologies du web, qui deviennent toujours plus performantes. Les portails web de certaines bibliothèques reposent encore sur des moteurs de recherche limités et des interfaces de navigation vieillissantes ayant une influence néfaste sur l'attrait que doivent susciter les institutions culturelles. Dans ce contexte, le texte de Roy Tennant intitulé « *MARC must die* » est souvent cité dans la communauté documentaire pour illustrer l'ensemble de ces problèmes [131].

Depuis deux décennies, la communauté documentaire argumente en faveur d'un abandon des formats de notices bibliographiques au profit des technologies du Web Sémantique. Différents travaux de recherche appuient cette volonté comme dans la thèse de Melhem [86] où différentes enquêtes ont démontré le besoin d'adopter de nouvelles technologies pour faciliter la réutilisation des métadonnées bibliographiques. D'autres travaux ont montré les bénéfices apportés par l'adoption de nouvelles ontologies sémantiques, en remplacement des formats de notices, pour mieux modéliser les connaissances bibliographiques [106, 153], proposer de nouvelles visualisations des données aux utilisateurs, [88] ou encore faciliter la recherche de contenus intellectuels [119]. La publication, en 1998, des principes FRBR, pour *Functional Requirements for Bibliographic Records*, a officialisé un mouvement international pour l'évolution des systèmes d'information bibliographiques [132]. En France, un groupe nommé *Transition Bibliographique*[1], appuyé notamment par la Bibliothèque Nationale de France, a été créé pour orchestrer ce mouvement à l'échelle nationale. Leur travail s'étend notamment sur les règles, modèles et vocabulaires qui doivent remplacer les formats de notices pour la gestion des métadonnées bibliographiques ainsi que sur la transformation des millions de notices existantes vers ces nouveaux modèles sémantiques.

1.1.1 Problématique générale

Dans la communauté documentaire différents travaux ont été menés pour étudier les étapes que doivent réaliser les institutions afin de permettre cette adoption des principes du Web Sémantique pour la gestion des ressources bibliographiques (*ex.*, en France [121]). Alemu *et al.*, en étudiant ces différents travaux, soulèvent l'enjeu crucial de la transformation des catalogues bibliographiques : « *Whilst the opportunities presented by Linked Data are evident, it is nonetheless crucial to recognise the challenges that libraries are confronting especially in view of the significant size of legacy metadata still locked in MARC formats* » [9]. Le coût d'un projet de transformation des nombreuses notices existantes pouvant être très élevé, les documentalistes doivent avoir recours à un processus informatique spécifique qui facilite l'exécution de cette tâche [1, 82, 23]. La figure 1.2 montre les grandes phases de ce processus de transformation d'un catalogue bibliographique. Nous discutons ci-après des enjeux qui sont soulevés par un tel processus.

Sur la figure 1.2, on observe que le processus de transformation est réalisé en deux étapes, (1) la migration des notices du catalogue vers un ensemble d'entités et de relations puis (2) l'enrichissement de ces dernières avec des sources externes. Les deux étapes reposent sur un ensemble de règles qui sont définies par les experts et documentalistes. Si la seconde étape, c'est à dire l'enrichissement d'entités structurées avec des sources du Web Sémantique, est très étudiée dans différentes communautés, la première étape de migration est spécifique au contexte bibliographique. En effet, Aalberg *et al.*, ont montré que les connaissances bibliographiques d'un catalogue étaient implicitement représentées dans les champs des notices le composant, nécessitant alors un processus spécifique d'interprétation de ces connaissances [5]. De plus, ils démontrent qu'une mauvaise interprétation des données peut entrainer une perte d'informations importantes lors du processus de transformation empêchant par la suite toute inférence des connaissances. Par ailleurs,

1. `https://www.transition-bibliographique.fr`

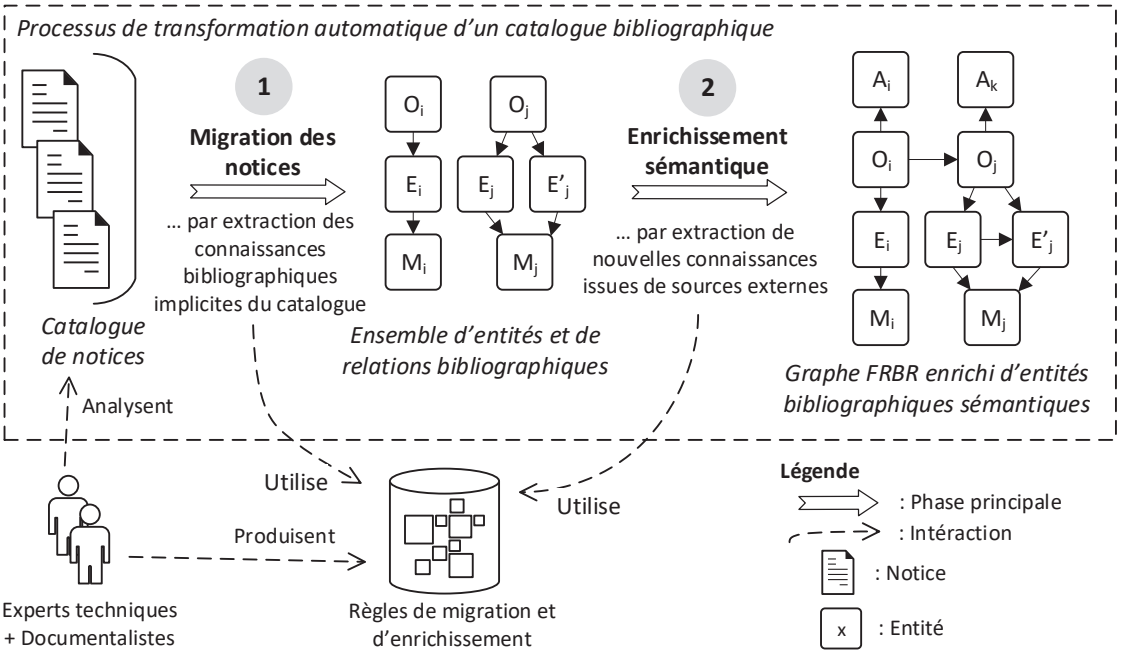


FIGURE 1.2 – Processus de transformation d'un catalogue de notices bibliographiques vers les principes du Web Sémantique

Baker confirme la nécessité de cette étape d'interprétation en soulevant l'importance d'un travail préalable d'harmonisation, par les institutions documentaires, des vocabulaires et modèles de données qui doivent être utilisés dans le Web Sémantique [14]. Enfin, d'autres travaux soulèvent les difficultés posées par la modélisation des connaissances qu'il est nécessaire de réaliser à la première étape d'un processus d'intégration de données culturelles [97, 7]. Cette modélisation est représentée, sur la figure 1.2, par une base de règles produite par les documentalistes. Si la communauté s'accorde sur la nécessité de transformer et enrichir les données bibliographiques, Hallo *et al.*, en dressant un état des lieux de l'utilisation des technologies du Web Sémantique dans les institutions documentaires, montrent que les systèmes d'information du domaine ne permettent pas encore de réaliser ces opérations [54]. Ils détaillent par exemple que des systèmes n'intègrent pas une modélisation des données selon le standard *Resource Description Framework*[2] (RDF) ou encore que certains systèmes ne référencent pas les entités par un mécanisme d'URIs[3] uniques facilitant l'alignement des données avec d'autres bases de connaissances.

A partir de ces observations nous présentons la problématique générale de cette thèse : **permettre la transformation des catalogues de ressources bibliographiques vers les technologies du Web Sémantique en respectant les contraintes de qualité des institutions documentaires**. Cette problématique inclut la capacité des outils de transformation à considérer les connaissances bibliographiques ainsi que le respect de la complétude et de la pertinence des données qui sont transformées, afin de permettre leur réutilisation dans le Web Sémantique.

2. https://www.w3.org/RDF/
3. http://www.bnf.fr/fr/professionnels/web_semantique_boite_outils/a.web_semantique_standards.html

1.1.2 Hypothèse principale

Les travaux que nous citons précédemment nous invitent à constater que les outils existants de migration de modèles de données, qui permettent de réaliser des alignements entre différents modèles (ici, entre un modèle de notices et une ontologie du Web Sémantique), n'intègrent pas les spécificités et le langage particulier du domaine bibliographique détaillés par Lubetzky en 1969 [81]. Cette limitation peut rendre plus difficile l'utilisation d'outils de migration par les institutions documentaires. Bernstein *et al.* explicitent ce problème : « *Given the existence of all these tools, why is it still so laborintensive to develop engineered mappings ? To some extent, it is an unavoidable consequence of ambiguity in the meaning of the data to be integrated. If there is a specification of the schemas, it often says little about integrity constraints, units of measure, data quality, intended usage, data lineage, etc.* » [17]. Les connaissances bibliographiques pouvant être riches et complexes, les schémas de règles (c-à-d., la structure des correspondances établies entre un modèle source et un modèle cible) utilisés par les outils de migration informatiques peuvent manquer de mécanismes spécifiques pour permettre la modélisation des connaissances bibliographiques de manière efficace et lisible. Bernstein *et al.* ajoutent : « *Since human designers are required, the solution must lie in raising the level of abstraction in which engineered mappings are specified and in offering better tools to do that specification* ». Les travaux présentés dans ce manuscrit suivent cette vision. Notre hypothèse principale est que la modélisation des règles de migration et d'enrichissement doit être adaptée au contexte spécifique des données bibliographiques. Cela doit permettre d'intégrer les experts du domaine documentaire dans le processus d'écriture de ces règles (et pas seulement des informaticiens) afin, d'une part, de faciliter les discussions autour des futurs modèles de données bibliographiques dans un contexte évolutif et, d'autre part, d'améliorer la qualité globale des outils de migration pour satisfaire les exigences du domaine culturel et ainsi accélérer la transformation des catalogues vers le Web Sémantique.

1.2 Synthèse des travaux réalisés

Les travaux réalisés dans le cadre de cette thèse s'intègrent dans un double contexte où, d'une part, la communauté documentaire a observé un besoin immédiat d'outils facilitant la migration des notices existantes vers le web de données et, d'autre part, les nouveaux modèles qui permettent la description des connaissances bibliographiques ne sont pas entièrement formalisés. Dis autrement, les critères de qualité et de pertinence des processus d'intégration de données bibliographiques ne sont pas largement unifiés car les travaux sur les modèles et vocabulaires associés (*ex.*, FRBR), dans le Web Sémantique, sont encore récents dans la communauté [90, 114].

1.2.1 Domaines de contributions

Pour répondre à ce double enjeu de transformation et de modélisation des métadonnées bibliographiques, nous adressons plusieurs défis scientifiques :

D'abord, nous nous intéressons à la mesure de la qualité des solutions de migration et d'enrichissement de notices bibliographiques. En suivant notre hypothèse principale, qui recommande une adaptation des règles de transformation au contexte bibliographique, l'évaluation de cette nouvelle catégorie d'outils nécessite des critères spécifiques au domaine. Nous définissons alors un ensemble de métriques qui tiennent compte des spécificités du domaine documentaire en terme de qualité des bases de connaissances bibliographiques. Ces métriques vont plus loin que les critères habituels de mesure de la qualité, comme la complétude des données, car elles considèrent les motifs de connaissances avancées qui sont essentielles au domaine, c'est à dire des combinaisons de relations sémantiques entre les données, qui forment la richesse bibliographique.

Nous adressons ensuite le problème de la modélisation des connaissances bibliographiques, avec les technologies du Web Sémantique, qui sont encore représentées implicitement dans les anciens formats de notices. Nous proposons une méthodologie originale permettant de créer un modèle de règles (pour la migration et l'enrichissement) qui tienne compte des spécificités du domaine bibliographique tout en favorisant la réutilisation de ces règles. Notre approche permet aux experts informatiques et documentalistes de facilement réaliser un tel processus de transformation en manipulant des concepts (c-à-d., connaissances bibliographiques) qui leurs sont familiers tout en respectant les exigences de qualité inhérentes aux données bibliographiques.

Nous répondons enfin au manque d'expérimentations et d'évaluations des processus de migration et d'enrichissement de catalogues bibliographiques dans la communauté documentaire. Nous proposons de nouveaux jeux de données qui intègrent une collection experte, réalisée manuellement, afin de mieux comparer les résultats produits par les différents outils. Nous proposons également une implémentation de notre modèle de règles nous permettant d'évaluer l'impact bénéfique de notre hypothèse principale dans le contexte de migrations de catalogues du monde réel.

1.2.2 Publications

Nous présentons ci-après la liste des publications principales issues de ce travail.

Revues internationales

— *(Article long)* Aalberg, T., Duchateau, F., Takhirov, N., Decourselle, Joffrey., & Lumineau, N. (2018, Janvier) Benchmarking and Evaluating the Interpretation of Bibliographic Records. International Journal on Digital Libraries (IJDL) (pp. 1-23) (SJR2016 : Q1).

— *(Article long)* Decourselle, J. (2016). Towards a Pattern-based Semantic Enrichment of Bibliographic Entities. IEEE Technical Committee on Digital Libraries (TCDL), 12(2).

Conférences internationales

— *(Article long)* Decourselle, J., Duchateau, F.,& Lumineau, N. (2015, Septembre). A Survey of FRBRization Techniques. International Conference on Theory and Practice of Digital Libraries (TPDL). Proceedings, Kapidakis S., Mazurek C., Werla M. Research and Advanced Technology for Digital Libraries. LNCS, vol 9316. (core2017 : A)

— *(Article long)* Decourselle J., Duchateau F., Aalberg T., Takhirov N., Lumineau N. (2016, Septembre) BIB-R : A Benchmark for the Interpretation of Bibliographic Records. International Conference on Theory and Practice of Digital Libraries (TPDL). Proceedings, Fuhr N., Kovács L., Risse T., Nejdl W. (eds) Research and Advanced Technology for Digital Libraries (pp. 163-174). (core2017 : A)

— *(Poster)* Decourselle, J., Duchateau, F., Aalberg, T., Takhirov, N., & Lumineau, N. (2016, Juin). Open Datasets for Evaluating the Interpretation of Bibliographic Records. IEEE/ACM Joint Conference on Digital Libraries (JCDL), Newark, NJ, (pp. 253-254). (core2017 : A*)

— *(Poster)* Decourselle, J. (2016, Septembre). Case-oriented Semantic Enrichment of Bibliographic Entities. International Conference on Theory and Practice of Digital Libraries (TPDL). (core2017 : A)

— *(Article industriel)* Decourselle, J., & Riondet, F. (2017, Octobre). Adopting Semantic Technologies in Public Health Documentation. In ISWC Industry Track. (core2017 : A)

Workshops

— *(Article long)* Decourselle, J., Vennesland, A., Aalberg, T., Duchateau, F., & Lumineau, N. (2015, Septembre). A Novel Vision for Navigation and Enrichment in Cultural Heritage Collections. East European Conference on Advances in Databases and Information Systems (ADBIS), Workshop Semantic Web For Cultural Heritage (SW4CH). Proceedings, New Trends in Databases and Information Systems (pp. 488-497).

Autres publications

Nous présentons une sélection de publications secondaires dans cette thèse :

— *(Poster)* Decourselle, J., Duchateau, F., & Ganier, R. (2016, Janvier). Syrtis : New Perspectives for Semantic Web Adoption. In BOBCATSSS.

— *(Article industriel)* Decourselle, J. (2017, Mai) Progilone : un acteur au cœur du renouveau des humanités numériques. Archimag

— *(Site Internet)* Prototype CoM3ET [Page web dédiée au prototype].
Disponible sur : `http://research.progilone.fr/com3t.html`

— *(Site Internet)* Benchmark BIB-R [Page web du benchmark].
Disponible sur : `http://bib-r.github.io/`

1.2.3 Applications industrielles

Cette thèse s'inscrit dans le cadre d'un contrat CIFRE entre l'entreprise Progilone [4] et le Laboratoire d'InfoRmatique en Image et Systèmes d'information [5] (LIRIS), tous deux basés à Lyon en France. L'objectif de ce partenariat est de répondre au besoin croissant, dans la communauté documentaire, de nouveaux systèmes permettant de gérer les métadonnées bibliographiques selon les principes du Web Sémantique et les nouvelles normes de catalogage *Resource Description and Access* [6] (RDA). La société Progilone, éditrice de solutions documentaires à Lyon, développe le logiciel Syrtis [7], qui permet de transformer des notices bibliographiques en une base de connaissances sémantiques, basée sur la norme RDA, puis de valoriser cette base avec diverses fonctionnalités d'enrichissement, de gestion et de diffusion des métadonnées. L'intégration de nouveaux catalogues dans la solution Syrtis exige un travail important de validation de la qualité et de la réutilisabilité des données qui sont intégrées dans le logiciel. Comme cette solution Syrtis se destine à la gestion de catalogues contenant de larges volumes de notices, l'automatisation des processus de migration et d'évaluation de la qualité des données est un enjeu industriel important pour l'entreprise Progilone. Les travaux réalisés dans le cadre de cette thèse ont donc été industrialisés afin d'être appliqués dans le contexte de la solution Syrtis de Progilone.

4. `https://www.progilone.fr`
5. `https://liris.cnrs.fr/`
6. `http://www.bnf.fr/fr/professionnels/normes_catalogage_francaises/a.rda_fr.html`
7. `https://www.progilone.fr/fr_FR/syrtis/`

1.3 Organisation du manuscrit

Ce manuscrit est découpé en 7 chapitres. Le présent chapitre définit la problématique générale de la thèse et présente l'hypothèse principale de recherche ainsi que les contributions proposées. Le chapitre 2 introduit certaines notions essentielles, issues du domaine bibliographique, qui sont requises pour la compréhension des contributions proposées. Le chapitre 3 décrit notre état de l'art sur l'intégration de données documentaires et sur la création de bases de connaissances bibliographiques. Le chapitre 4 présente et détaille l'ensemble des métriques de qualité permettant d'évaluer les étapes majeures de la transformation des catalogues documentaires, en bases de connaissances sémantiques, et propose un benchmark original permettant d'évaluer les solutions d'interprétation de métadonnées bibliographiques. Le chapitre 5 décrit notre méthodologie pour construire un modèle de migration de métadonnées bibliographiques qui tienne compte des exigences de qualité définies dans le chapitre précédent afin de permettre la réalisation d'un processus de migration et d'enrichissement de catalogues bibliographiques. Le chapitre 6 présente l'outil COM3ET (intégré au logiciel Syrtis) qui consiste en une implémentation des approches présentées dans les chapitres précédents, notamment au niveau de la migration des catalogues documentaires, et présente des expérimentations permettant de valider notre hypothèse principale. La conclusion et les perspectives de cette thèse constituent le chapitre 7 de ce manuscrit.

Chapitre 2

Prérequis

Ce chapitre décrit le contexte de la gestion de données bibliographiques dans la communauté documentaire. Nous introduisons les notions essentielles à l'étude des processus de migration et d'enrichissement sémantique des métadonnées bibliographiques.

Sommaire

2.1 Introduction

La communauté documentaire rassemble l'ensemble des acteurs et des solutions qui gravitent autour des objectifs communs de création, préservation et valorisation des ressources documentaires des domaines privés comme publiques. Le travail documentaire respecte des normes qui ont évolué au rythme des révolutions technologiques et sociétales. Avec l'émergence du Web Sémantique, notre rapport à l'information a évolué, notamment aux niveaux de la recherche et de la visualisation des informations. Dans le contexte bibliographique, ces changements, amenés par les technologies numériques incitent les documentalistes à repenser la manière dont ils gèrent les métadonnées qui décrivent les ressources documentaires. L'objectif principal pour ces professionnels est de faciliter l'accès au patrimoine culturel et intellectuel pour les nouvelles générations d'utilisateurs en tenant compte des possibilités technologiques, comme le Web Sémantique.

Les travaux qui sont présentés dans cette thèse s'inscrivent dans l'élan d'évolution des systèmes d'informations documentaires vers le Web Sémantique. Afin d'appréhender les spécificités et enjeux de cette évolution, nous commençons par présenter les modèles et pratiques bibliographiques actuels puis nous détaillons les méthodes et les solutions, qui doivent être progressivement adoptées par la communauté, pour la création des futures bases de connaissances sémantiques.

2.2 Description des connaissances bibliographiques

Les catalogues documentaires représentent la source d'information principale des différents processus présentés dans cette thèse. Ces catalogues permettent la description de chaque ressource documentaire d'une institution. Ces ressources peuvent être des livres, des articles, des textes officiels, des périodiques, ou encore du multimédia. Les éléments permettant la description de ces ressources sont appelés les métadonnées bibliographiques. Dans la suite, nous étudions les spécificités de ces métadonnées d'un point de vue intellectuel puis structurel. Plus précisément, nous examinons les connaissances qui sont représentées dans les métadonnées, puis nous détaillons la manière dont ces connaissances sont structurées dans les modèles bibliographiques existants.

2.2.1 Connaissances sur les ressources documentaires

Les catalogues auxquels nous nous intéressons sont composés de notices décrivant les ressources (documents) que possède une institution documentaire. Leazer et Smiraglia ont décrit, en 1999, le concept (encore largement utilisé) de "*notice*" comme une instance des connaissances que l'on possède sur un document : « *A bibliographic entity is a unique instance of recorded knowledge [...] [and] has both physical and intellectuel properties* » [79]. Aujourd'hui, les notices numériques remplacent les anciennes notices au format papier. Nous nous interessons au contenu intellectuel des notices avant de présenter, toujours dans cette section, la structure d'une notice.

Les connaissances présentes dans un catalogue bibliographique sont à la fois des informations éditoriales (*ex.*, ISBN, dates) et des informations intellectuelles sur les documents d'une institution. Les informations éditoriales, servant à identifier et localiser un document (physiquement dans une institution), sont dépendantes des standards du domaine et des pratiques métier (*ex.*, côtes des livres, nomenclatures spécialisées). Dans le contexte de cette thèse, nous nous intéressons plus particulièrement aux connaissances intellectuelles d'une ressource. Ces dernières, qui ont été longuement étudiées par la communauté (*ex.*, travaux de Tillet [133]), détaillent les informations sur la création originale d'un document (titres, résumés, langues) ainsi que sur les différentes entités associées à cette création (auteurs, concepts, ressources liées). Une création originale et ses entités liées forment une famille bibliographique, c'est à dire un sous-ensemble de l'écosystème des métadonnées bibliographiques d'une institution. La Figure 2.1 présente un exemple d'écosystème bibliographique dans une institution documentaire.

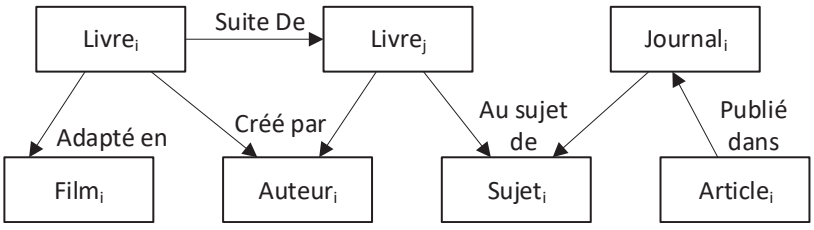


FIGURE 2.1 – Exemple de l'écosystème bibliographique d'une institution documentaire

Dans la terminologie documentaire, les connaissances intellectuelles sur les ressources bibliographiques (*ex.*, Livre, Film, Article ou Journal dans la figure 2.1) sont symbolisées par des *œuvres* intellectuelles qui sont liées à d'autres œuvres ou d'autres entités (*ex.*, sujet d'une œuvre) pour former des familles bibliographiques [122] (où chaque œuvre décrit sa propre famille).

L'œuvre bibliographique

Les connaissances bibliographiques, que l'on souhaite réutiliser, sont des ensembles d'entités et de relations qui s'articulent autour des créations intellectuelles qui composent un catalogue. Par exemple, *Les Liaisons dangereuses* de *Pierre Choderlos de Laclos* est une création intellectuelle qui a donné lieu à de nombreuses éditions (*ex.*, traductions) et adaptations (*ex.*, films ou pièces de théâtre). Comme explicité dans les travaux de Carlyle [24], chaque création intellectuelle est appelée une *Œuvre* dont la description est issue de l'agrégation d'une ou plusieurs notices.

Définition 2.2.1 (Œuvre bibliographique)**.** La notion d'Œuvre ("Work" en anglais) est issue d'une proposition de Lubetzky, en 1969, d'agréger des notices dont le contenu intellectuel est identique (*ex.*, les différentes notices décrivant les livres ou films issues des *Liaisons dangereuses*) [81]. L'objectif de ce regroupement est de simplifier la recherche bibliographique dans un catalogue en limitant le nombre de résultats aux créations originales plutôt qu'aux éditions. *Svenonius*, en 2000, a proposé la définition suivante : « *A Work is a set or family of documents in which each document embodies essentially the same information or shares essentially the same intellectual or artistic content* » [123]. Cette définition a été complétée, par *Smiraglia*, pour considérer l'Œuvre comme une entité, à un niveau plus abstrait, permettant la description d'une création originale : « *A work is a signifying, concrete set of ideational conceptions that finds realization through semantic or symbolic expression. That is, a work embraces a set of ideas that constitute both the conceptual (signified) and image (signifier) components of a sign.* » [122].

Pour résumer, un catalogue bibliographique est composé de notices. Chaque notice décrit une ressource documentaire de l'institution en charge du catalogue. Plus particulièrement, chaque ressource est une édition spécifique issue d'une création originale et intellectuelle, l'œuvre. Ainsi, chaque ressource de l'institution, décrite par une notice est issue d'une œuvre. C'est pourquoi plusieurs notices d'un catalogue peuvent faire référence à une même œuvre (chaque notice représente alors une édition de cette œuvre). La Figure 2.2 présente la distinction entre la création intellectuelle (l'œuvre) abstraite et les éditions qui sont décrites par les notices d'un catalogue.

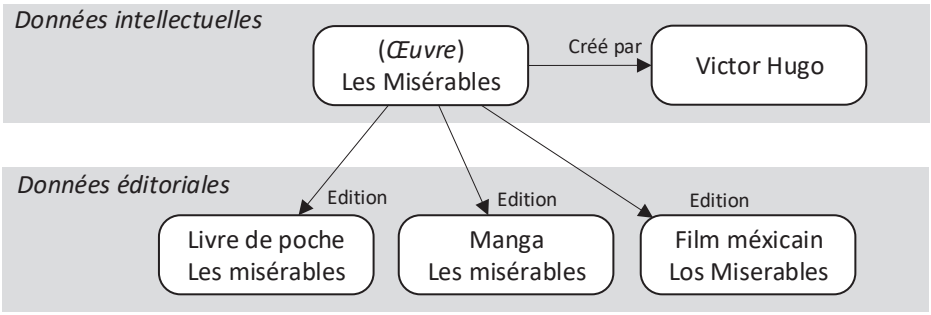


FIGURE 2.2 – Notices d'un catalogue issues d'une même création intellectuelle

Famille bibliographique d'une œuvre

Le point central des connaissances d'un catalogue repose sur la notion d'œuvre bibliographique. C'est au travers de cette notion qu'un utilisateur peut explorer les différentes réalisations et dérivations de cette œuvre. Dans l'exemple de la Figure 2.2, ces informations sont implicitement décrites entre la création intellectuelle et les différentes éditions. Les variations intellectuelles d'une œuvre s'incarnent par un ensemble de relations, issues de l'œuvre en question, vers d'autres entités du catalogue ou d'autres sources. Vellucci, en 1997 [137], a nommé cet ensemble de relations comme la *famille bibliographique* d'une œuvre.

Définition 2.2.2 (Famille bibliographique)**.** Une famille bibliographique est un ensemble de relations organisées autour d'une œuvre (appelé aussi *progenitor*), qui sert de point d'entrée pour une exploration intellectuelle d'un contexte bibliographique (*ex.*, adaptations ou éditions de l'oeuvre). Leazer et Smiraglia, en 1999, ont proposé la description suivante : « *Any work, either a wholly new work or a work derived from some other work, can serve as a progenitor for additional works. [...] The set of such interrelated works and items is called a bibliographic family, where the relationships among entities express shared semantic or linguisitc activity, [...]. bibliographic relations would assist users in selecting the most suitable work for their own purposes.* » [79].

Exemple 2.2.1. Harry Potter Le domaine de fiction *Harry Potter* est à l'origine d'un large ensemble de ressources bibliographiques comme des livres et des films. Dans ce domaine, différentes familles bibliographiques peuvent être observées. Si l'on étudie la famille bibliographique issue de l'œuvre principale, c'est à dire le premier livre de *J.K. Rowling* (Harry Potter and the Philosopher's Stone), on pourrait observer l'adaptation du roman en film, réalisé par *David Yates*, ou encore les multiples traductions du livre qui ont été réalisées. Une autre famille bibliographique peut être constituée à partir du concept abstrait de Harry Potter comme "super"-œuvre à partir de laquelle on pourrait observer l'ensemble des œuvres, c'est à dire les romans et films associés.

2.2.2 Taxonomie des relations bibliographiques

Nous présentons maintenant les caractéristiques des familles bibliographiques. Ces dernières ont été la source d'études, dans la communauté documentaire, afin d'améliorer les systèmes dédiés à la navigation et à la représentation des ressources documentaires. Des travaux significatifs, réalisés notamment par Tillett, Smiraglia ou Vellucci ont abouti à la proposition de taxonomies et classifications des relations bibliographiques qui composent ces familles (*ex.*, les *adaptations*, *critiques* ou *suites* d'une œuvre). Noruzi a présenté une synthèse de ces travaux dans [94]. Les relations bibliographiques ont donc été répertoriées et classifiées. Par ailleurs, une pratique que nous suivons dans cette thèse consiste à réutiliser ces classifications comme base de modélisation des connaissances bibliographiques. Nous présentons ici les trois principales catégories de relations que nous considérons : les *dérivations*, les *agrégations* et les *œuvres connexes*. Nous omettons les relations qui sont spécifiques à la publication éditoriale des ressources (*ex.*, *réimpressions*, *facsimilés*) et qui ne concernent pas directement le contexte intellectuel de l'œuvre en question. Nous présentons, ci-après, ces différentes catégories de relations bibliographiques.

Définition 2.2.3 (Dérivations)**.** Les dérivations d'œuvres font référence à toutes les modifications intellectuelles réalisées sur le contenu original d'une œuvre donnée. Les traductions, les adaptations, les abréviations ou les révisions sont des dérivations d'œuvres.

Exemple 2.2.2. L'adaptation en film, avec *Audrey Tautou*, du roman *The Da Vinci Code* est une dérivation de l'œuvre originale de *Dan Brown*.

Il existe de nombreuses dérivations possibles d'une œuvre. Aussi, Smiraglia a proposé une sous-classification des types de dérivations [94] :

- Les *dérivations successives ou simultanées* (*ex.*, les œuvres révisées plusieurs fois de suite).
- Les *traductions* (*ex.*, les œuvres traduites dans différentes langues).
- Les *augmentations* (*ex.*, les œuvres augmentées par des illustrations ou des préfaces).
- Les *adaptations* (*ex.*, œuvres basées sur une œuvre originale, dans un format différent).
- Les *réductions* (*ex.*, les extraits ou résumés d'une œuvre).
- Les *performances* (*ex.*, les réalisations ponctuelles d'une œuvre).

Nous notons que, dans les travaux présentés plus tard dans ce manuscrit, nous considérons *les augmentations* comme une catégorie de relations à part entière. En effet, la modélisation des augmentations dans une base de connaissances peut varier selon l'importance que l'on accorde à ces augmentations dans un catalogue de ressources bibliographiques donné. Par exemple, une augmentation apportée à une œuvre littéraire, comme une illustration, peut être considérée comme une œuvre à part entière ou non, ce qui influence la manière dont elle doit être modélisée.

Définition 2.2.4 (Agrégations)**.** Les agrégations concernent tous types de liens créés entre des œuvres ou des éditions d'œuvres servant à former des ensembles et groupes d'informations ayant un intérêt intellectuel ou éditorial. Nous considérons deux principales catégories d'agrégations qui sont les *relations de tout ou parties* et *les relations séquentielles*. Dans le premier cas, il s'agit de relations établies entre la partie d'une création intellectuelle et sa totalité (*ex.*, la trilogie du *Seigneur des anneaux*). Dans le second cas, il s'agit des relations établies entre une œuvre ou une édition d'œuvre qui en précède une autre (*ex.*, les différents fascicules d'un journal périodique).

Exemple 2.2.3. Un recueil de poésie, une collection jeunesse dans une bibliothèque ou encore les différentes suites d'un livre ou d'un film constituent différentes formes d'agrégations possibles.

Définition 2.2.5 (Œuvres connexes)**.** Les œuvres connexes sont liées par des relations spécifiques qui ne sont ni des dérivations, ni des agrégations. Ces relations sont créées pour faciliter l'exploration d'une famille bibliographique au travers de liens définis entre différentes ressources. Les œuvres connexes peuvent être matérialisées par des relations entre une œuvre principale et d'autres œuvres qui décrivent, critiques ou évaluent l'œuvre principale.

Exemple 2.2.4. L'œuvre, *Le Da Vinci Code expliqué à ses lecteurs* de *Bernard Sesboüé*, est une œuvre connexe à l'œuvre *The Da Vinci Code* de *Dan Brown* par une relation d'*analyse*. D'autres liens peuvent être créés par les institutions, à destination des usagers, pour lier entre-elles des œuvres, comme des suppléments ou des œuvres partageant des caractéristiques communes.

Nous avons présentés les principales catégories de relations bibliographiques que nous incluons dans la notion de *connaissances bibliographiques*. Pour rappel, un objectif des différents travaux présentés dans cette thèse consiste à valoriser ces connaissances présentés précédemment dans les futures bases de connaissances sémantiques des institutions documentaires.

2.2.3 Principes de catalogage

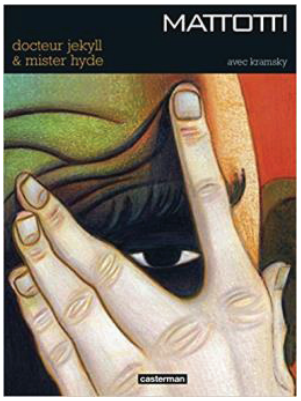
Nous avons présenté les contenus intellectuels des catalogues de ressources bibliographiques. Dans cette partie, nous présentons comment ces informations intellectuelles sont organisées et stockées dans le modèles de données d'une institution documentaire. Nous parlons ici de la tâche de *catalogage*, réalisée par les professionnels documentaires. Par ailleurs, nous étudions comment les relations bibliographiques entre les œuvres sont représentées dans ces modèles de catalogage.

Modèle de catalogage

Pour rappel, chaque ressource possédée par une institution documentaire est décrite par une notice. Chaque notice est composée de champs sous la forme de paires "clé/valeur". Chaque clé est un code correspondant à une sémantique selon une spécification et la valeur associée à une clé contient une information sur la ressource qui est décrite dans la notice. La Figure 2.3 présente un exemple de notice bibliographique, avec ses champs, décrivant une bande dessinée. Dans cet exemple, le titre de la ressource, *Docteur Jekyll & mister Hyde*, est stocké sous la clé *200$a*.

Cette notice utilise un format de type MARC [1], pour *MAchine-Readable Cataloging*. Ce type de format, en liste de champs (*ex.*, *200$a*), date des années 60 et a été conçu durant l'effort de

1. `http://www.enssib.fr/le-dictionnaire/marc-formats`



010 $a 2-203-38988-5
101 $a fre $c it
200 $a Docteur Jekyll & mister Hyde / dessin, Lorenzo Mattotti
$b Texte imprimé
$e adapt. d'après l'oeuvre originale de Robert L. Stevenson
210 $a [Paris] $c Casterman $d 2002
215 $a 64 p. $c couv. ill. en coul., $d 32 cm
606 $a Littératures $a Classique
700 $a Mattotti $b Lorenzo $4 070
702 $a Voline $b Marc $4 730
700 $a Stevenson $b Robert Louis $4 100
995 $p Disponible

FIGURE 2.3 – Notice bibliographique au format UNIMARC

numérisation des anciennes notices, initialement au format papier. Au moment de la rédaction de ce manuscrit, les formats MARC sont encore les plus répandus dans le monde, pour le catalogage de données bibliographiques. Toutefois, l'aspect *MAchine-Readable* de ces formats n'est plus d'actualité et doit évoluer vers les technologies du Web Sémantique. Dans le contexte de cette thèse, les travaux ont été principalement réalisés avec le format UNIMARC, utilisé très largement en France et dans d'autres pays d'Europe (*ex.*, Espagne, Portugal, Italie). La notice de la Figure 2.3 est donc présentée via ce format. Toutefois, les principes et propriétés que nous illustrons dans cette thèse s'appliquent à tous les types de formats dérivés des principes MARC.

Les notices étant la source d'information principale pour la migration et l'enrichissement des catalogues bibliographiques, la tâche initiale de ces processus consiste à interpréter ces notices.

Définition 2.2.6 (Interprétation d'une notice)**.** L'interprétation d'une notice consiste en l'analyse intellectuelle de cette dernière afin d'identifier les connaissances utiles qu'elle décrit. De manière plus schématique, une tâche d'interprétation prend en entrée une notice bibliographique et produit un ensemble de fonctions qui, à partir de combinaisons spécifiques de données de la notice, retournent une connaissance utile du domaine bibliographique.

Exemple 2.2.5. Nous illustrons cette tâche d'interprétation avec l'exemple de la Figure 2.3. Ici, une interprétation correcte de la notice est : *une bande dessinée, réalisée en italien par Lorenzo Mattotti, puis traduite en Français par Marc Voline, qui est une adaptation de Docteur Jekyll & mister Hyde, créée initialement par Robert L. Stevenson.* Il s'agit donc d'une œuvre qui a été adaptée en bande dessinée, elle-même réalisée en italien puis traduite en français. Dans cette notice, chaque champ correspond à l'association d'un code de zone à trois chiffres (*ex.*, 010, 100, 200, ...) et d'un caractère (lettre ou chiffre). Le tout est joint par le caractère '$'. Ainsi, 200$*a*, 210$*d* ou encore 700$*b* sont des champs auxquels est associée une sémantique. Par exemple, 200$*a* correspond au titre de publication, 210$*d* correspond à la date de publication et 700$*b* est le prénom d'une responsabilité principale de la publication, selon la spécification UNIMARC.

Connaissances intellectuelles implicites en MARC

Dans un format MARC, l'organisation des informations sur une ressource est avant tout pensée pour la tâche de catalogage (donc de stockage) et pas directement pour la restitution des données aux utilisateurs. Ainsi, les informations intellectuelles (c-à-d., les connaissances) sur une ressource peuvent être dispersées dans plusieurs champs issus de différentes zones. Dans l'exemple de la Figure 2.3, les informations concernant la traduction en français de la bande dessinée sont séparées entre les langues dans la zone 101 et le traducteur en 702$4. Également, les informations relatives à l'adaptation en bande dessinée sont séparées entre les champs 200$*a* et $*e* (description),

les champs de la zone 215 (formats) et les champs 700$4 (rôles). De plus, le niveau sémantique des champs et des zones n'est pas nécessairement homogène. En effet, une entité à part entière (*ex.*, une personne, un lieu), peut être décrite par une zone (*ex.*, zones 700 ou 702 faisant référence à des responsabilités de l'œuvre) ou par un champ (*ex.*, champs 210$*c* qui décrit un éditeur).

Pour résumer, une notice peut décrire plusieurs entités qui composent la famille bibliographique d'une œuvre. Plusieurs notices peuvent être également nécessaires pour reconstituer la famille dans son intégralité. Il est important de noter que les relations qui constituent une famille bibliographique sont généralement décrites de manière implicite dans les notices. Par exemple, la notice de la Figure 2.3 contient deux traductions et une adaptation de l'œuvre de *Stevenson* qui ne sont pas forcément déductibles de manière triviale. Les travaux présentés dans cette thèse s'intéressent aux efforts d'interprétation et de modélisation des connaissances issues des notices.

Pratiques de catalogage de notices

L'analyse des modèles de catalogage de notices, qui sont largement utilisés, nous montre comment les informations intellectuelles sont implicitement représentées, impliquant ainsi des efforts importants pour les interpréter. Dans cette partie, nous décrivons comment cette difficulté d'interprétation est accentuée par les pratiques spécifiques de catalogage. En effet, les systèmes d'information documentaires sont souvent basés sur des versions adaptées des spécifications bibliographiques. Cela veut dire que les règles de catalogage des ressources bibliographiques sont adaptées aux contraintes spécifiques d'une institution documentaire. Par exemple, un catalogue bibliographique, sur la recherche médicale, n'utilisera pas les mêmes champs qu'un catalogue de lecture publique. Le domaine médical aura besoin d'adapter les règles de catalogage à la modélisations d'auteurs scientifiques ayant des affiliations. L'interprétation des relations bibliographiques dans les notices peut donc varier d'un catalogue à l'autre en fonction de ces spécificités. Nous considérons deux principaux types d'adaptation, au niveau *modèle* et au niveau *instance*.

Les *adaptations au niveau du modèle* correspondent à des ajouts de champs spécifiques ou à toutes les modifications, au niveau des cardinalités des champs, qui ne sont pas conformes à la spécification utilisée (*ex.*, UNIMARC). Dans l'exemple de la Figure 2.3, le champ 955$*p* a été créé spécifiquement pour gérer la disponibilité de la bande dessinée dans la bibliothèque. De plus, la zone 606, normalement prévue pour ne décrire qu'un seul concept, contient ici deux concepts (Littérature et Classique) dans chaque champ $*a*. Cette pratique est régulièrement employée pour intégrer des valeurs de concepts contrôlées issues de thésaurus spécifiques (*ex.*, Rameau [2]).

Les *adaptations au niveau des instances* concernent les modifications faites sur la sémantique initialement prévue par la spécification. Ces modifications peuvent être réalisées par une extension de la spécification standard ou par l'ajout de codes spécifiques, placés directement dans les valeurs des champs. Dans l'exemple de la Figure 2.3, le champ 200$*a*, initialement prévu pour contenir le titre de la ressource, contient ici plus d'informations pour faciliter un affichage spécifique de type ISBD [3]. De plus, le champ 200$*e* contient une information sur l'adaptation de l'œuvre, mais en langue naturelle. Ces pratiques sont utilisées pour répondre à des contraintes d'affichage ou logicielles. Par ailleurs, elles peuvent également être employées pour spécialiser la sémantique d'un champ, issue de la spécification, mais considérées comme trop vague par une institution.

2. `http://rameau.bnf.fr/informations/rameauenbref.htm`
3. `http://bnf.fr/fr/professionnels/normes_catalogage_intles/a.normes_isbd_presentation.html`

2.3 Modélisation des métadonnées bibliographiques

Dans la section précédente, nous avons présenté la façon dont les informations documentaires d'une institution sont cataloguées dans des notices bibliographiques. Dans le contexte de cette thèse, nous nous intéressons à la transition de ces informations, issues des notices, vers des modèles et systèmes plus en phase avec les principes du Web Sémantique. Ces nouveaux principes, présentés dans [9], [33] ou encore [54] intègrent progressivement les notions d'œuvre et de familles bibliographiques de manière explicite. Ainsi, la communauté tend à favoriser la valorisation des connaissances intellectuelles dans les futures pratiques et systèmes de catalogage. Dans cette section, nous présentons ces nouveaux principes et technologies ainsi que les enjeux associés.

2.3.1 Functional Requirements for Bibliographic Records (FRBR)

À partir des travaux réalisés, à la fois sur les œuvres et familles bibliographiques issues des catalogues (voir section précédente), ainsi que sur l'évolution des technologies de l'information, de nouvelles propositions de modélisation des métadonnées documentaires ont été étudiées.

Origine et principes de FRBR

Dans le contexte bibliographique, les principes FRBR [132] nés au début des années 1990, se sont imposés progressivement comme remplaçant des modèles MARC [131, 85]. FRBR, pour *Functional Requirements for Bibliographic Records*, consiste en une méthodologie conceptuelle pour la représentation de métadonnées bibliographiques qui est radicalement différente de la modélisation historique MARC [60, 148, 31]. Si dans les formats de type MARC les métadonnées éditoriales et intellectuelles de chaque ressource documentaire sont mélangées et "aplaties" sur une notice composée de paires clé-valeur, la modélisation selon FRBR est un graphe d'entités et d'associations sémantiques inspiré du web de données. Un des changements principaux est que FRBR préconise que les données bibliographiques soient désormais centrées sur l'œuvre intellectuelle à partir de laquelle peuvent être déclinées des relations à plusieurs niveaux du contexte intellectuel vers le contexte éditorial [16, 49]. FRBR permet donc de représenter de manière explicite les familles bibliographiques des œuvres et de mieux distinguer les différents niveaux d'information (intellectuel et/ou éditorial) de chaque œuvre d'une institution [60, 154].

De nombreux travaux ont été menés par la communauté documentaire pour appréhender ces nouveaux principes FRBR et étudier la possibilité de créer de nouveaux modèles de bases de données bibliographiques qui intègreraient ces principes [60, 154, 31]. Des expériences d'implémentation de FRBR ont été réalisées dans le monde entier comme en Autriche [108], en Australie [11], aux États-Unis [13] ou encore en Égypte [40]. Les recherches se sont ensuite focalisées sur les efforts d'adaptation nécessaires pour passer des modèles historiques, comme MARC21 [112], UNIMARC [104], ou plus globalement la famille MARC [2] aux principes FRBR.

Ces différentes expérimentations ont donné lieu à divers ajustements et compléments des principes FRBR, comme par exemple pour la gestion des publications en série (*ex.*, périodiques) [77], des agrégations d'œuvres [155] ou encore des concepts sujets des œuvres [52]. Ces évolutions ont mené à la publication, en 2015, des principes FRBR/LRM [114] (où LRM signifie *Library Reference Model*) désignés comme nouvelle version standard de FRBR. Les travaux de cette thèse ont systématiquement intégré les évolutions des principes FRBR. Aussi, pour éviter toute confusion, **nous considérons que les appellations "FRBR", "FRBR/LRM" et "LRM" qui sont utilisées dans ce manuscrit ainsi que dans les références citées désignent toutes la version la plus récente de FRBR** dont nous décrivons les aspects techniques ci-après.

Description technique de FRBR

FRBR, à l'inverse de MARC, ne propose pas de solution figée et "prête à l'emploi" pour la modélisation des connaissances documentaires mais plutôt un modèle conceptuel et un ensemble de recommandations pour construire une ontologie de description bibliographique qui puisse être implémentée dans un système d'information documentaire. Ce modèle conceptuel permet donc de guider la modélisation des futures bases de connaissances. Il est structuré en trois groupes qui sont respectivement, la *description bibliographique*, les *responsabilités* et les *concepts associés*. Chaque groupe contient un ensemble de classes abstraites qui doivent servir de base à la réalisation d'un modèle de données bibliographiques. La Figure 2.4 présente une vue schématique du modèle conceptuel FRBR selon ces trois groupes.

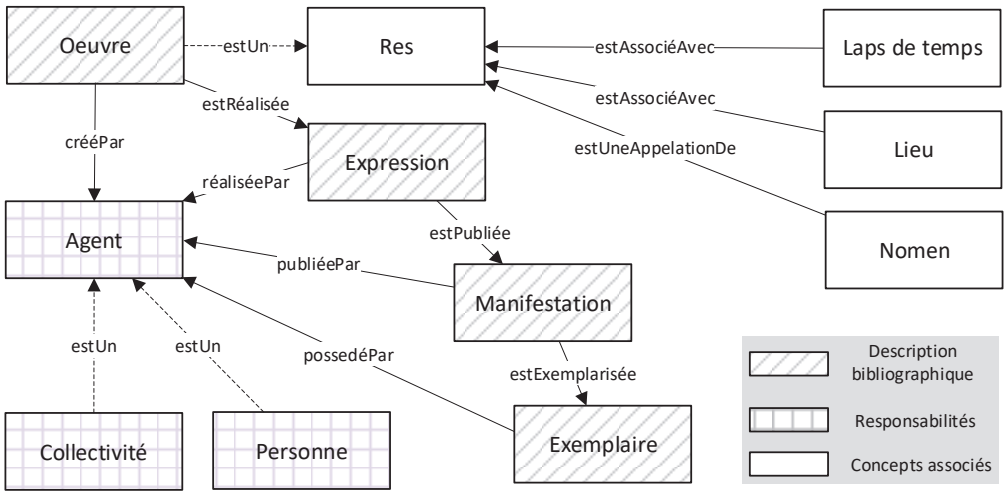


FIGURE 2.4 – Modèle conceptuel FRBR

Le premier groupe de FRBR concerne la **description bibliographique** et est divisé en quatre niveaux sémantiques, symbolisés par quatre classes Œuvre, Expression, Manifestation et Exemplaire. Ces quatre niveaux permettent la description d'une création intellectuelle, de ses réalisations dans des langues ou supports différents ainsi que ses différentes publications qui sont conservées et mises à disposition par des institutions. L'*Œuvre* représente la création intellectuelle et originale, telle que nous l'avons défini précédemment dans ce chapitre. L'*Expression* décrit toutes les réalisations intellectuelles d'une *Œuvre* comme les traductions ou les illustrations. La *Manifestation* est la publication d'une ou plusieurs *Expressions* dans un format physique ou électronique. Enfin, l'*Exemplaire* contient les informations permettant la localisation ou la gestion d'un objet physique d'une *Manifestation* qui serait possédée par une institution.

Le deuxième groupe de FRBR permet de décrire les **responsabilités**, c'est à dire les personnes ou collectivités qui interviennent dans la création, réalisation, publication ou possession d'une œuvre ou de ses éditions. Dans FRBR, une responsabilité est, de manière abstraite, symbolisée par la classe *Agent*. Cette dernière peut être spécialisée en une classe *Personne* ou une classe *Collectivité* selon les besoins de modélisation.

Le troisième groupe de FRBR concerne les **concepts associés** qui servent à compléter les informations des classes des deux précédents groupes. Dans FRBR, toutes les classes de ces groupes sont des sous-classes de *Res*. Par exemple, une *Œuvre* ou un *Agent* sont des sous-classes de *Res*. Pour des raisons de lisibilité, la Figure 2.4 ne montre qu'un seul lien de subsomption "*estUn*" avec *Res*. Chaque classe issue de Res peut-être donc liée aux classes *Lieu*, *Laps de Temps* et *Nomen*. Un *Lieu* peut être le sujet d'une *Œuvre* ou encore l'emplacement de naissance d'un *Agent*. Un *Laps de Temps* peut décrire les dates de vie d'une Personne (Agent) ou

encore une date de publication d'une *Manifestation*. La classe *Nomen* permet de gérer les appellations des différentes classes comme des points d'accès contrôlés, comme détaillés dans [103]. Par exemple, dans une modélisation utilisant RDF, l'URI des entités issues des classes de FRBR (*ex.*, http ://.../agent#victorhugo) est alors considérée comme un point d'accès contrôlé qui permet de lier la ressource à l'ensemble de ses appellations (*ex.*, *Victor-Marie Hugo*). Autre exemple, le journal *L'Obs* peut être modélisé par une entité de classe Œuvre (sous-classe de *Res*) qui serait elle-même liée à plusieurs appellations (= Nomen) comme *L'Obs* ou encore *Le Nouvel Observateur*. En pratique, ces appellations (aussi appelées *autorités*) sont représentées et gérées dans une base annexe comme un thésaurus.

2.3.2 Vers une évolution des pratiques de gestion bibliographique

Les modèles de données bibliographiques évoluant (*ex.*, avec l'apparition de FRBR), le contenu même des métadonnées bibliographiques évolue aussi avec l'utilisation grandissante des technologies numériques (*ex.*, jeux vidéo, ebooks, podcasts, sites web) pour la diffusion de contenus intellectuels. Ces évolutions ont un impact sur les règles de catalogage, c'est à dire les règles qui définissent la manière dont les informations bibliographiques doivent être stockées, ainsi que sur la manière de concevoir et maintenir les systèmes d'informations documentaires. Coyle et Hillmann ont résumé ces changements en présentant l'avenir des règles de catalogage avec l'arrivée des principes FRBR [32]. Nous dressons ci-après un état des lieux synthétique des modèles et règles qui sont expérimentés et progressivement adoptés par la communauté documentaire, dans le contexte de cette thèse. La Figure 2.5 propose une chronologie de l'évolution des modèles et règles de catalogage jusqu'au moment le plus récent que nous pouvons présenter. Pour des raisons de lisibilité, nous ne présentons que les éléments ayant un impact majeur, à notre sens, sur les communautés bibliographiques et musées.

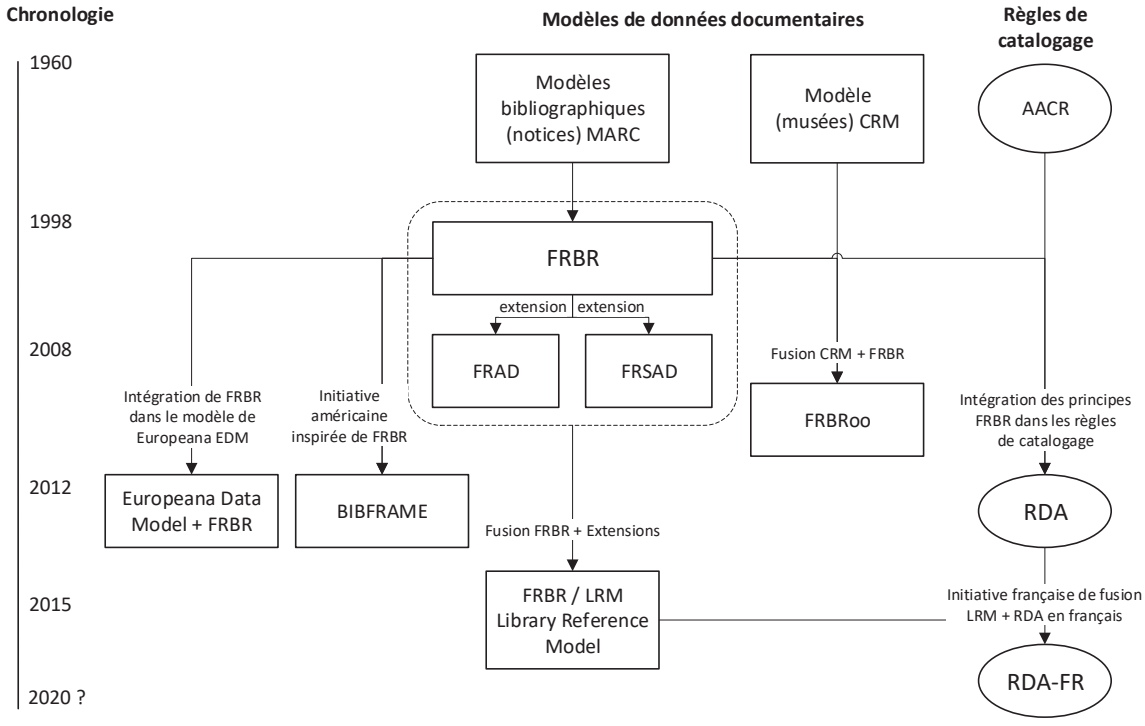


FIGURE 2.5 – Évolution des modèles et règles de catalogage

Les modèles de données privilégiés

Les modèles de données documentaires, présentés dans la Figure 2.5, ont été répertoriés et étudiés par Coyle [31] et plus récemment par Zapounidou *et al.* [150]. Les projets majeurs, pour l'adoption du Web Sémantique dans les institutions documentaires, s'articulent autour des modèles (plus ou moins conceptuels) *Europeana Data Model (EDM)*, *BIBFRAME*, *FRBRoo* et *FRBR*. EDM, propulsé par la plateforme numérique et européenne de description bibliographique, Europeana [4], a été établi avant FRBR et EDM mais s'est plus tard enrichi d'une partie de ces nouveaux principes. BIBFRAME est une initiative propulsée par la Bibliothèque du Congrès américaine qui reprend globalement le principe d'Œuvre, comme proposé initialement par FRBR, qu'elle fusionne avec la notion d'Expression pour proposer une version simplifiée des principes FRBR. Ce compromis s'explique par le fait que le format MARC21, utilisé très largement par les anglo-saxons, est la version de MARC la plus éloignée des principes FRBR, rendant ainsi leur adoption plus complexe. La simplification des principes FRBR dans BIBFRAME permet une adoption plus en douceur de ces nouveaux principes par les utilisateurs de MARC21. Les communautés documentaires plus spécifiques, comme les musées ou encore celles en charge du patrimoine musical, s'orientent plus largement vers FRBRoo. Ce dernier, inspiré à l'origine du modèle de données des musées, *CRM* [5], vient enrichir les principes FRBR de notions plus avancées sur la temporalité de création d'une œuvre, qui sont des notions essentielles pour ces communautés. Enfin, d'autres projets d'évolution bibliographique qui ne sont pas directement liés à Europeana ou à la bibliothèque du congrès, utilisent des versions adaptées de FRBR. En France, le projet *Data.BNF* de la bibliothèque nationale de France expérimente une base de connaissances sémantiques et bibliographiques qui s'inspire des principes FRBR [120]. Le projet français *Doremus* complète le projet *Data.BNF* sur le patrimoine musical en se basant sur le modèle FRBRoo [6].

De nouvelles règles de catalogage adoptées

Les modèles MARC s'accompagnent de règles de catalogage qui définissent la manière de gérer les informations dans les champs des notices. Les règles *Anglo-American Cataloguing Rules* [6] (AACR) sont les plus utilisées dans la communauté documentaire. Elles régissent par exemple la cardinalité des champs, l'ordre des informations ou encore les données obligatoires selon le type de ressource décrite. Avec l'émergence de nouveaux modèles, issues de FRBR, de nouvelles règles ont été proposées pour être plus en phase avec ces principes, notamment la notion d'Œuvre intellectuelle. Ces règles, appelées RDA pour *Resource Description and Access*, sont reconnues comme la référence pour le catalogage bibliographique et comme successeur des règles AACR. RDA intègre les niveaux sémantiques de FRBR (*ex.*, Oeuvre, Expression, Manifestation) et prévoie une gestion des données bibliographiques en phase avec les standards du Web Sémantique. Toutefois, RDA est encore en cours de finalisation à l'échelle internationale et devrait obtenir une nouvelle version en 2019. En France, un groupe de travail prépare une version de RDA adaptée aux spécificités françaises, appelé RDA-FR. L'adoption de RDA-FR dans les bibliothèques françaises n'est toutefois pas attendue avant au moins 2020. Quoi qu'il en soit, la ligne directrice de la France, présentée par le groupe français de Transition Bibliographique à la fin 2017, officialise l'abandon des modèles et règles liés aux anciennes notices bibliographiques pour une adoption de FRBR/LRM et RDA-FR.

4. https://www.europeana.eu/
5. http://www.bnf.fr/fr/professionnels/modelisation_ontologies/a.modele_cidoc_crm.html
6. https://en.wikipedia.org/wiki/Anglo-American_Cataloguing_Rules

2.4 Conclusion

Dans ce chapitre, nous avons présenté les spécificités des ressources bibliographiques qui représentent la source de données principale des processus détaillés dans cette thèse. Nous avons introduit, d'une part, l'existant en termes de modèles de données des catalogues bibliographiques et, d'autre part, les nouveaux modèles sémantiques qui sont au cœur de l'évolution des systèmes documentaires. Les critiques des modèles de notices que l'on observe dans la communauté révèlent qu'un travail important d'interprétation des connaissances bibliographiques existantes est nécessaire en amont de la transformation des modèles de données vers le Web Sémantique. De plus, il est important de noter que les nouveaux standards de modèles bibliographiques, comme FRBR, sont des recommandations qui impliquent une mûre réflexion sur la modélisation et l'instanciation des métadonnées afin de valoriser ces dernières dans les futures bases de connaissances.

Les outils de migration et d'enrichissement de métadonnées doivent tenir compte du changement de paradigme impliqué par la nouvelle modélisation des connaissances bibliographiques. Il est donc essentiel de connaître les possibilités des solutions existantes de transformation des métadonnées et d'identifier les nouveaux enjeux soulevés par cette évolution dans les modèles bibliographiques, principalement en termes de qualité des futures bases de connaissances. Cette étude sur les principaux outils existants et leurs objectifs est l'objet du prochain chapitre.

Chapitre 3

État de l'art

Dans ce chapitre, nous étudions l'état de l'art des solutions de migration et d'enrichissement de métadonnées pour construire des bases de connaissances sémantiques. En accord avec les objectifs de cette thèse, nous nous focalisons sur la migration de métadonnées bibliographiques.

Sommaire

3.1 Introduction

Les travaux réalisés dans cette thèse s'inscrivent dans le contexte de la valorisation des ressources documentaires des institutions culturelles qui sont dépendantes d'un système d'information numérique. Plus particulièrement, nous étudions les processus de migration et d'enrichissement appliqués aux données provenant de ces systèmes. Dans cette introduction, nous motivons notre choix d'étudier le contexte et les perspectives de ces deux processus et nous dressons une vue d'ensemble des travaux liés à ce domaine d'étude.

Une étude sur les caractéristiques des systèmes d'informations numériques (aussi nommés "Digital Libraries" ou *DL*) a été réalisée par Fuhr *et al.* [51]. Ils analysent les différentes facettes de ces systèmes comme les types de données qui sont gérés (*ex.*, contenu des ressources, métadonnées) et les fonctionnalités et services qui peuvent être appliqués sur ces types de données (*ex.*, recherche, sélection, mise en relation de documents). Les travaux menés notamment par Coyle et Merčun étudient eux les leviers possibles de valorisation des ressources culturelles en tenant compte de ces systèmes d'information documentaires [30, 89]. Dans nos travaux, nous nous focalisons sur deux leviers de valorisation : la modélisation des métadonnées bibliographiques selon les principes FRBR et l'enrichissement des métadonnées bibliographiques avec des sources externes.

Le premier levier de valorisation est la modélisation des métadonnées bibliographiques selon les principes FRBR (*cf.*, prérequis développés dans le chapitre précédent). Cette modélisation, étudiée dans les travaux de Taylor [129] ou de Pisanski *et al.*, [107] apporte de multiples bénéfices pour les systèmes d'informations documentaires. Nous avons vu dans le chapitre 2 que FRBR permet une recherche intuitive de documents (via la notion d'Œuvre) et une représentation enrichie des liens entre les entités (*ex.*, adaptations, traductions, augmentations) [88]. Toutefois, les travaux de Aalberg *et al.* [5] ou de Pattuelli et Cristina [102] montrent les difficultés d'implémenter ce type de modélisation dans les systèmes d'informations culturels existants, qui reposent sur d'anciens formats comme MARC. Nous détaillons ces problèmes d'implémentation (plus loin) dans ce chapitre. Le second levier de valorisation concerne l'enrichissement des métadonnées bibliographiques avec des sources externes comme le Web Sémantique. Dans les études de Pandey *et al.* [100] ou Hallo *et al.* [54] on observe les bénéfices apportés par l'ajout de nouvelles connaissances dans les catalogues bibliographiques via des sources externes. Cependant, les deux études dressent un constat similaire quant aux limitations des systèmes existants pour intégrer ces connaissances. Afin d'appréhender les défis inhérents à ces deux leviers de valorisation, nos travaux se focalisent sur la migration des métadonnées bibliographiques vers les principes FRBR et l'enrichissement de ces dernières avec des sources de données externes.

3.1.1 Périmètre d'étude

Les systèmes d'informations documentaires que nous considérons contiennent des métadonnées bibliographiques qui sont gérées selon des formats non FRBR (*ex.*, MARC) et qui sont isolées des référentiels externes comme le web de données ou les catalogues nationaux (*ex.*, BNF en France). Les processus de migration et d'enrichissement de ces métadonnées ont donc pour objectif l'évolution des systèmes d'informations documentaires pour mieux valoriser les ressources qu'ils contiennent. La figure 3.1 présente l'ensemble des phases relatives aux deux processus.

Sur la figure 3.1, le processus de migration comporte les phases de *transformation des notices* et de *déduplication*[1] (le terme déduplication est fréquemment utilisé dans la littérature de la communauté [49, 5] et est, dans ce contexte, équivalent à la tache de Record Linkage[2]). Le processus d'enrichissement comporte les phases d'*alignement d'entités*, d'*extraction de données* et également de *déduplication*. Ces différents processus sont successivement détaillés dans la suite de ce chapitre. Nous faisons remarquer que les lettres utilisées pour représenter les entités dans nos exemples d'illustration (ex., O_i) désignent plus spécifiquement des instances de classes du modèle conceptuel FRBR (A = Agent, O = Œuvre, E = Expression et M = Manifestation). Par ailleurs, dans le reste de ce manuscrit, l'utilisation d'une première lettre majuscule appliquée à ces termes (ex., Œuvre, Expression) fait systématiquement références aux classes de FRBR.

Dans cette illustration de la figure 3.1, la transformation du catalogue local donne lieu à trois Œuvres O_i, O_j et O_k avec leurs autres entités associées. La premières déduplication fusionne O_j et O_k qui sont en réalité la même œuvre. De plus O_i et O_j sont créées par le même Agent A_j (*ex.*, un même auteur ayant écrit deux livres). Lors de l'enrichissement, l'entité O_j sera alignée avec une entité d'une source externe, permettant d'extraire une autre Œuvre associée O_n. (*ex.*, O_n est la suite de O_j). Enfin, la deuxième déduplication révèle que O_n est créée par le même Agent A_i que O_i et O_j (*ex.*, c'est le même auteur qui a écrit les trois Œuvres O_i, O_j et O_n). Les deux processus de migration et d'enrichissement sont réalisés de manière consécutive. En effet, d'après les travaux sur l'interprétation des métadonnées bibliographiques réalisés notamment par Hickey *et al.* [60], Riva *et al.* [112], ou encore Freire *et al.* [49], la phase de transformation des notices doit inclure une étape d'interprétation des données, selon les connaissances bibliographiques for-

1. https://en.wikipedia.org/wiki/Data_deduplication
2. https://en.wikipedia.org/wiki/Record_linkage

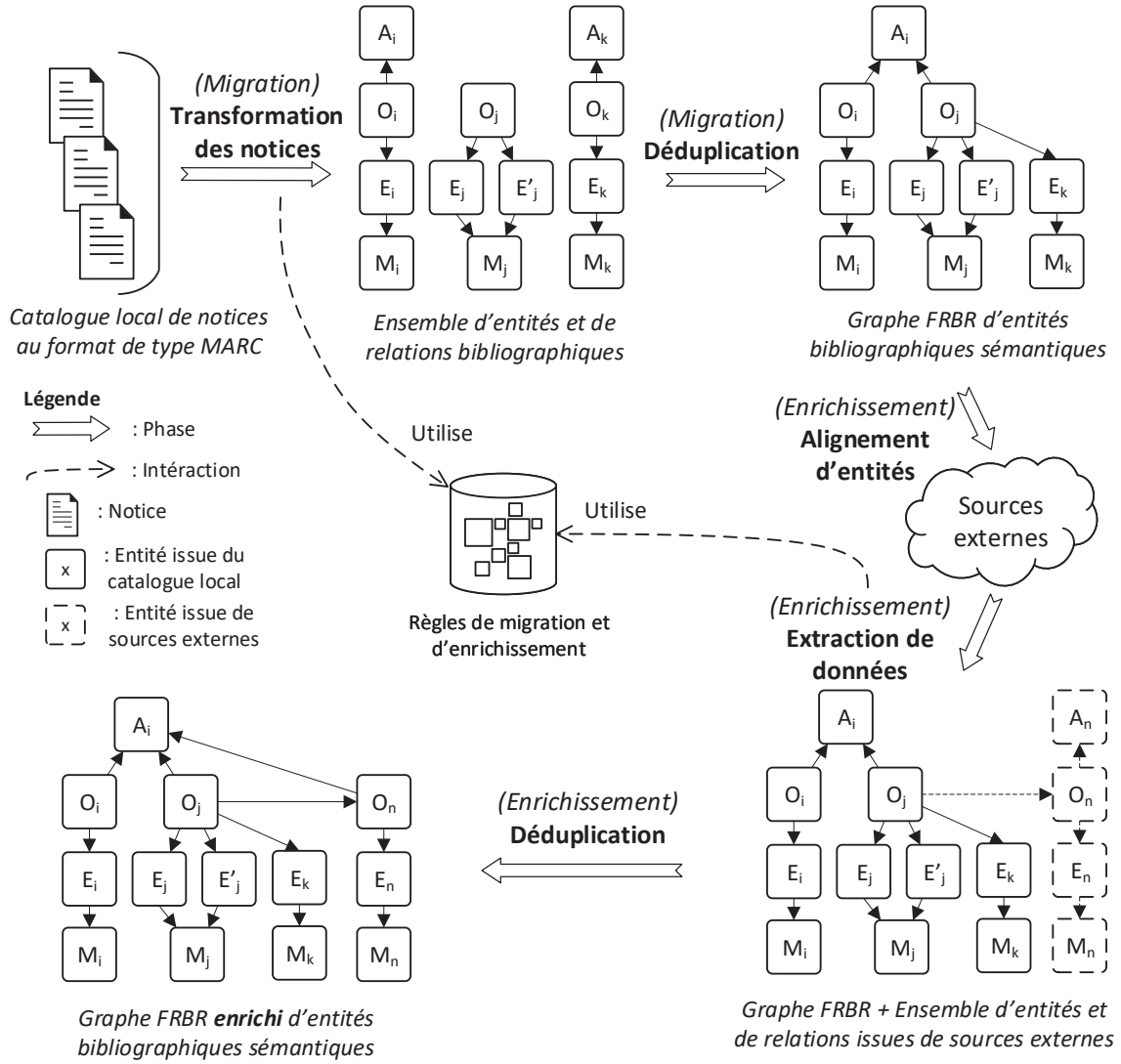


FIGURE 3.1 – Vue d'ensemble des phases de migration et d'enrichissement de métadonnées bibliographiques dans l'objectif de valorisation des ressources documentaires.

malisées par la communauté [133], afin de pouvoir exploiter les relations FRBR produites. Par exemple, la quantité d'informations nécessaires à l'alignement de O_j avec des sources externes peut être dépendante de l'interprétation réalisée du catalogue et de la première déduplication entre O_j et O_k. C'est pourquoi, dans le contexte des métadonnées documentaires, l'enrichissement sémantique n'est réalisé qu'une fois qu'un processus de migration vers FRBR est terminé.

Dans ce chapitre, nous étudions les approches existantes concernant les phases qui sont présentées dans la figure 3.1. La section 3.2 décrit les méthodes de transformation des notices bibliographiques vers les principes FRBR, la section 3.3 étudie le processus de déduplication dans ce contexte et la section 3.4 présente les méthodes facilitant l'interopérabilité des entités bibliographiques avec des sources de données externes. La section 3.5 dresse un bilan de l'état de l'art présenté et motive les contributions de cette thèse.

3.2 Transformation des notices bibliographiques

L'adoption des principes FRBR est un levier pour une meilleure valorisation des ressources bibliographiques. FRBR, impliquant notamment la modélisation des métadonnées selon un graphe d'entités et d'associations, est cependant très éloigné des formats existants de notices (*cf.*, chapitre 2) et implique une transformation importante dans les métadonnées des catalogues existants. En ce sens, bien qu'une partie de la communauté ait préconisé l'abandon complet des formats de notices au profit de FRBR [131, 110], toutes les institutions ne sont pas encore équipées pour réaliser immédiatement cette transformation. Aalberg *et al.* ont proposé une approche intermédiaire qui décrit des méthodes de nettoyage et de réorganisation des champs des notices pour préparer une future transition vers FRBR, et bénéficier partiellement de ces principes, tout en conservant un format de type MARC [2]. Une partie des institutions peut cependant décider d'abandonner complètement les formats de notices et doivent entament une transformation de leurs catalogues existants et des outils permettant de gérer leurs métadonnées. Dans l'optique d'un abandon complet des formats de type MARC, la phase de transformation des catalogues sous-entend la migration des notices existantes vers les principes FRBR, appelé aussi processus de *FRBRisation* [37]. Ce processus essentiel pour l'adoption des principes du web de données dans les institutions documentaires est au cœur des travaux et réalisations de cette thèse.

3.2.1 Techniques de FRBRisation

Nous présentons deux techniques de FRBRisation qui sont majoritairement utilisées dans la communauté : le *regroupement des notices* et l'*extraction des connaissances*.

Regroupement des notices

Le regroupement des notices est une technique spécifique au domaine bibliographique qui se concentre sur le niveau Œuvre de FRBR. Pour rappel, le niveau Œuvre correspond à la création intellectuelle à partir de laquelle plusieurs éditions concrètes peuvent avoir été publiées (niveau Manifestation FRBR). Le choix est fait dans cette méthode de ne pas considérer le niveau intermédiaire d'Expression FRBR. Nous précisons ce choix ci-après. Les notices étant associées à une publication, elles représentent le niveau Manifestation. Ainsi, le regroupement de notices doit permettre de détecter différentes Manifestations dans le catalogue issues d'une même Œuvre. Comme présenté par Dickey, l'avantage de ce regroupement est de réduire le nombre de résultats (Œuvres) affichés à un utilisateur qui cherche dans un catalogue documentaire [37]. La technique de regroupement repose sur la détection de similarités entre les notices d'une même Œuvre. Dit autrement, il s'agit d'un processus de détection d'entités similaires dont l'objectif est d'isoler certaines données communes à un niveau d'abstraction plus élevé. Hickey *et al.* ont été parmi les premiers à expérimenter cette technique [60] sur un échantillon du catalogue WorldCat [3]. Leurs résultats ont montré que certaines combinaisons élémentaires de critères sur les notices (*ex.*, *titre*, *auteur principal*) suffisait à obtenir de bons résultats dans la qualité du regroupement au niveau Œuvre. Cependant, il a été montré qu'il est beaucoup plus difficile de trouver de bons critères pour la détection des Expressions (niveau intermédiaire). Le regroupement de notice a également été expérimenté par Hegna *et al.* [58] et Juffinger *et al.* [69]. Ces travaux ont obtenu des résultats similaires et ont choisi de ne pas considérer le niveau Expression de FRBR au profit de l'agrégation des notices au niveau Œuvre. Hickey a par ailleurs proposé, à partir des expérimentations réalisées, un algorithme (*WorkSet Algorithm*) pour la génération de clés permettant la comparaison de notices afin de détecter efficacement des Œuvres avec cette technique [61].

3. http ://0-www.worldcat.org.novacat.nova.edu/

Cette méthode et cet algorithme d'une faible complexité présentent l'avantage d'offrir un bénéfice immédiat et important de la modélisation FRBR, à savoir une simplification des résultats de recherche documentaire en réduisant le nombre d'éléments affichés à l'utilisateur (à savoir des Œuvres uniques). De plus, cette méthode permet d'évaluer la présence de doublons dans une collection d'Œuvres déjà transformées vers FRBR. Cependant, cette technique n'exploite pas le niveau Expression de FRBR car, selon les observations de Hickey, une détection correcte des Expressions nécessite une interprétation plus fine des notices [60]. Cette limitation peut être rédhibitoire pour certains projets de valorisation de ressources documentaires car le niveau Expression, qui a été largement étudié par Tillet, permet notamment de distinguer les types de contenus disponibles pour une même Œuvre (*ex.*, audio, texte numérique, etc.), les langues traduites ou encore les particularités des Œuvres comme les augmentations ou les révisions apportées au contenu [132]. Il peut être donc nécessaire pour une institution d'intégrer des niveaux de description intellectuelle pour les métadonnées des ressources qui soient plus fins que le niveau Œuvre, et donc nécessiter des efforts importants pour adapter la technique de regroupement des notices. Nous n'utilisons pas cette technique dans nos travaux pour cette même raison.

Extraction des connaissances

L'extraction des connaissances est la seconde technique de FRBRisation qui s'inspire des processus de transformation d'un modèle source (ici MARC) vers un modèle cible (ici FRBR) en utilisant un ensemble de règles prédéfinies [73]. Ces règles permettent d'associer des éléments du modèle source (*ex.*, une combinaison de champs de notice) avec des éléments du modèle cible (ex., deux entités avec une relation sémantique) pour ensuite migrer les données automatiquement d'un modèle vers un autre. Dans différentes communautés, ce type de processus, également associé au domaine du *Data Exchange* [59], a été largement appliqué au transfert de données issues d'anciens modèles vers des ontologies du Web Sémantique comme dans le domaine médical [56] ou le domaine universitaire [115]. Fagin *et al.*, dans le cadre du projet Clio, détaillent la nature et la structure des règles de transformation d'un point de vue global [45]. Ils montrent notamment que les règles sont construites à différents niveaux sémantiques comme des correspondances entre les éléments des modèles (*ex.*, classes ou propriétés) et des associations plus complexes entre des motifs au niveau *méta* de ces modèles (relations entre méta-modèles).

Au delà de la structure des règles, le contexte des métadonnées bibliographiques implique des recommandations particulières pour les règles de FRBRisation. En effet, ces dernières doivent associer des valeurs textuelles (MARC) à des graphes d'entités et d'associations (FRBR) comme par exemple la présence d'un titre de collection dans une notice d'un ouvrage menant à la création d'une relation sémantique entre deux entités FRBR de classe Œuvre, l'ouvrage et la collection. Walkowska et Werla dressent un ensemble de prérequis pour la création de règles de transformation d'un modèle comme des fonctions d'interprétation des valeurs dans les champs des notices, la création d'URIs uniques pour la construction des entités du modèle cible ou encore l'utilisation d'*Ontology Paths* (voir [96]) pour modéliser la partie droite des associations entre MARC et FRBR [140]. D'autres recommandations peuvent s'appliquer à des contextes plus spécifiques comme les données multimédia [64]. Les expérimentations du projet de FRBRisation TelPlus dressent également un ensemble de prérequis pour la création d'un modèle de règles de FRBRisation en fonction des caractéristiques du catalogue à transformer en entrée du processus [82].

Aalberg a implémenté le premier processus de FRBRisation basé sur un ensemble de règles de transformation de modèles, appelé *marc2frbr* [1]. Il décompose le processus en quatre étapes : (1) l'identification des entités, (2) l'assignation des attributs, (3) la détection des relations entre entités et (4) la normalisation des résultats. Les étapes (1) et (2) sont assurées par les correspondances qui sont définies entre les notices et les éléments du modèle cible. L'étape (3)

consiste à enrichir ces correspondances de fonctions et conditions plus spécifiques pour préciser la sémantique des relations entre les entités FRBR. Par exemple, une correspondance entre la responsabilité principale d'une notice et l'Agent d'une Œuvre en FRBR peut être enrichie d'une fonction analysant le code de fonction de la responsabilité dans la notice afin de préciser le rôle de l'Agent en FRBR (*ex.*, auteur, narrateur). L'étape (4) implique essentiellement des mécanismes d'intégration de données dont la déduplication des entités produites à l'étape (1). Le processus de déduplication est l'objet de la prochaine section de ce chapitre.

Ces deux techniques de FRBRisation (regroupement de notices, l'extraction de connaissances) permettent de produire un nouveau catalogue de métadonnées qui respecte partiellement ou complètement les principes FRBR. Si le regroupement de notices ne considère pas le niveau Expression de FRBR, l'extraction de connaissances implique une plus grande complexité dans l'écriture des règles de transformation des notices. Comme précisé précédemment, nous nous focalisons sur la (deuxième) technique d'extraction des connaissances pour son potentiel d'exploitation de l'ensemble des principes et éléments du modèle FRBR.

3.2.2 Outils de FRBRisation

Nous étudions maintenant les outils existants de FRBRisation, c'est à dire les solutions qui implémentent l'une ou l'autre des deux techniques présentées précédemment. Les outils de FRBRisation ont été listés et présentés dans différentes études réalisées par la communauté documentaire [152, 130, 37, 5]. Nous proposons notre propre étude de ces outils qui considère des approches plus récentes et encore non répertoriées et qui intègre une classification originale des solutions. Si notre sélection de solutions de FRBRisation peut ne pas être exhaustive avec le temps, elle révèle cependant un ensemble de caractéristiques du domaine couvrant un large ensemble de critères permettant de distinguer les approches et de discuter des forces et faiblesses de ces dernières.

Parmi les outils que nous présentons, certains sont des solutions commerciales et d'autres sont des projets de recherche. Les solutions **PRIMO** de ExLibris [108], **Classify** de OCLC [139] et **Virtua** de VTLS [43] sont des solutions commerciales qui offrent des fonctionnalités pour la gestion d'un catalogue bibliographique ainsi que des fonctions spécifiques pour l'adoption plus ou moins avancée des principes FRBR. Les trois solutions ne proposent pas une transformation du modèle de données bibliographiques mais permettent de classifier et d'agréger les notices selon les principes FRBR, en conservant un modèle de stockage MARC, afin d'améliorer les interfaces de recherche et de navigation. PRIMO et Classify réalisent cette classification en utilisant une technique de regroupement des notices inspirée du *WorkSet Algorithm*. Virtua propose la même fonctionnalité mais va plus loin en permettant de paramétrer les règles de classification des notices selon les principes FRBR. Le détail de cette fonctionnalité n'est cependant pas rapporté dans la littérature. Néanmoins, nous pouvons considérer que Virtua se rapproche plus de la seconde technique de FRBRisation même si la classification est réalisée "à la volée" et que les principes FRBR ne sont pas implémentés directement dans la gestion du catalogue.

Les autres solutions de FRBRisation sont des outils disponibles librement qui sont tous basés sur la seconde technique de FRBRisation d'extraction des connaissances avec des règles. Parmi les solutions existantes, certaines prennent la forme de logiciels comme les applications **Extensible Catalog** (XC) [20] et **Variations/FRBR** (VFRBR) [111] qui sont deux applications de FRBRisation codées en Java. XC, qui est une solution complète de gestion de métadonnées bibliographiques, est composée d'un module (appelé *Metadata Service Toolkit*) qui permet de récolter des notices basées sur des formats MARC, de transformer ces données vers FRBR et de normaliser les entités FRBR produites, notamment en détectant les doublons. VFRBR est une application dédiée à la FRBRisation de notices de documents musicaux, dont le catalogue

Scherzo [95] est un résultat concret de son utilisation en conditions réelles. L'inconvénient de ces deux solutions XC et VFRBR est que les règles de migration sont codées en dur dans l'outil, nécessitant des efforts importants pour adapter ces outils à différents contextes.

D'autres solutions résolvent ce problème en proposant un mécanisme de règles déclaratives, c'est à dire indépendantes de l'algorithme de transformation des notices. C'est le cas de l'outil **LC Display Tool** qui permet de transformer des notices utilisant un format MARC/XML grâce à des règles XSL [116]. Display Tool traite chaque notice indépendamment des autres et, pour chaque transformation effectuée, retourne une seule Œuvre FRBR, liée à une Expression, elle-même liée à une seule Manifestation. L'outil étant plutôt dédié à une utilisation pédagogique de FRBR, les entités retournés n'exploitent qu'une partie limitée des principes de FRBR (*ex.*, pas de relations bibliographiques avancées entre les trois entités proposées). **FRBR-ML**, version la plus récente de l'outil marc2frbr [1], est une solution proposant un langage de règles, au format XML, qui permet de gérer des aspects plus complexes de l'extraction des connaissances vers FRBR [127]. Notamment, FRBR-ML repose sur des règles orientées entités et contextualisées. Cela veut dire qu'il est possible d'assigner des conditions particulières à une classe de FRBR (*ex.*, Œuvre, Expression, Manifestation) afin de produire des entités FRBR qui sont issues de différents contextes bibliographiques (*ex.*, une Œuvre issue du titre originale dans le cas d'une traduction ou du titre propre dans un cas basique).

Nous illustrons cette notion de règles orientées entités avec le code XML de la figure 3.2 qui est tiré du modèle de règles proposé avec FRBR-ML. Il décrit une règle orientée-entité permettant la détection d'une entité de classe Expression dans le contexte d'une notice au format MARC21 qui contiendrait une zone au code 130. Les propriétés de l'entité Expression, comme `title` ou `languageOfExpression` sont encapsulées dans cette règle. En bas du code sont représentées les relations avec d'autres entités contextualisées (ici une entité de classe Manifestation). Cette structuration des correspondances de migration est fortement couplée au modèle MARC21 car les codes de ce format sont directement ajoutés dans la règles (il n'y a pas de modèle pivot intermédiaire). Ce choix s'explique car l'outil utilise des processus XSLT pour directement transformer des notices codées en MARC21/XML. L'autre problème vient de l'encapsulation des propriétés d'une entité à ce niveau (*ex.*, `contentType` qui est inclus dans cette règle). La création d'une autre règle pour décrire une Expression dans un autre contexte impliquera de dupliquer les propriétés dans cette nouvelle règles (*ex.*, redéfinir la propriété `contentType` à chaque fois). En conséquence, la description de motifs de connaissances issus de multiples contextes bibliographiques peut introduire beaucoup de redondance dans les règles avec un modèle orienté entité comme FRBR-ML, ce qui réduit la lisibilité du modèle et augmente les risques d'erreurs.

La solution **X3ML** propose un modèle de règles XML différent qui offre la possibilité de créer des règles, non pas centrées sur une entité particulière, mais sur un sous-ensemble du modèle cible [83]. De manière plus technique, la partie gauche d'une correspondance dans cet outil est une relation du modèle sources quand la partie droite peut être un graphe d'entités et d'associations décrivant plusieurs relations sémantiques. L'illustration de la figure 3.3 présente un exemple de modélisation d'une règle avec X3ML. Dans cet exemple, la propriété *weights* entre les classes Coin et Weight du modèle source doit produire un graphe de 5 classes et 4 relations dans le modèle cible. Cette méthode permet d'intégrer des motifs de connaissances (modèle cible) qui sont plus riches que dans le cas de règles orientées entités et qui correspondent mieux aux relations bibliographiques avancées (*cf.*, chapitre 2). X3ML ne propose cependant pas de règles prédéfinies pour le domaine bibliographique et n'a pas encore été expérimentée dans ce contexte.

La possibilité pour les experts de modifier les règles de FRBRisation et de bénéficier de mécanismes avancés pour gérer des cas de modélisation FRBR complexes dans les règles permet

```
<entity type="c:Expression" templatename="MARC21-130-Expression">
  <note>Expression of the work identified in field 130</note>
  <anchor tag="130"/>
  <attributes>
    <datafield tag="245">
      <subfield code="a" type="u:title"/>
    </datafield>
    <datafield tag="041">
      <subfield code="a" type="e:languageOfExpression"/>
      <subfield code="d" type="e:languageOfExpression"/>
      <subfield code="e" type="e:languageOfExpression"/>
      <subfield code="f" type="e:languageOfExpression"/>
      <subfield code="j" type="e:languageOfExpression"/>
    </datafield>
    <datafield tag="130">
      <subfield code="l" type="e:languageOfExpression"/>
    </datafield>
    <datafield tag="336">
      <subfield code="a" type="e:contentType"/>
    </datafield>
  </attributes>
  <relationships>
    <relationship type="e:manifestationOfExpression" itype="m:expressionManifested">
      <target entity="MARC21-001-Manifestation"/>
    </relationship>
  </relationships>
</entity>
```

FIGURE 3.2 – Extrait du modèle de règles de FRBR-ML

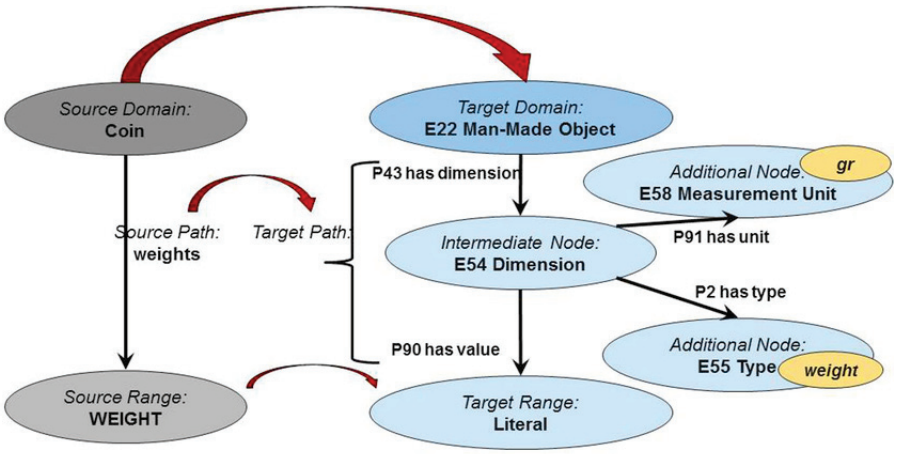


FIGURE 3.3 – Illustration de règles de migration dans l'outil X3ML [83].

d'améliorer la qualité d'un processus de FRBRisation. Cependant, dans le cas de projets de FRBRisation de catalogues très spécifiques et hétérogènes, l'écriture au format XML des règles et la validation des correspondances peut être une tâche très fastidieuse. De manière plus globale, les modèles de règles reposant sur un format nécessitant une expertise informatique restent difficiles à manipuler par les documentalistes dont le vocabulaire de travail et les enjeux pour un projet de migration sont très différents de ceux des informaticiens.

Certaines solutions de transformation de modèles proposent des interfaces graphiques de gestion des correspondances de transformation afin d'impliquer les documentalistes dans le processus d'écriture des règles. L'outil **LibFRBR** propose une interface utilisateur permettant de consul-

ter et modifier ces correspondances [25]. Cet outil, initialement prévu pour la FRBRisation de notices du patrimoine chinois (format CMARC) et développé en Perl, ne fournit cependant que peu d'informations sur son fonctionnement. La solution **Karma Tools** s'inscrit dans cette catégorie d'outils qui proposent de simplifier l'écriture des correspondances de transformation [71]. Karma intègre des mécanismes similaires à ceux de X3ML pour gérer des correspondances vers des motifs riches dans le modèle cible. La force de Karma réside dans sa possibilité d'afficher l'ensemble des règles de transformation de manière graphique afin de permettre aux experts du domaine (et non informaticiens) d'évaluer les correspondances et donc d'anticiper des erreurs d'interprétation en amont du traitement des notices. L'illustration de la figure 3.4 présente un exemple de modélisation de règles de migration dans l'outil Karma. Les champs en bleu (*ex.*, UID, job_title, position_type, etc.) sont issus du modèle source quand les éléments du graphe au dessus (*ex.*, *Position1 relates FacultyMember1*) sont issus du modèle cible.

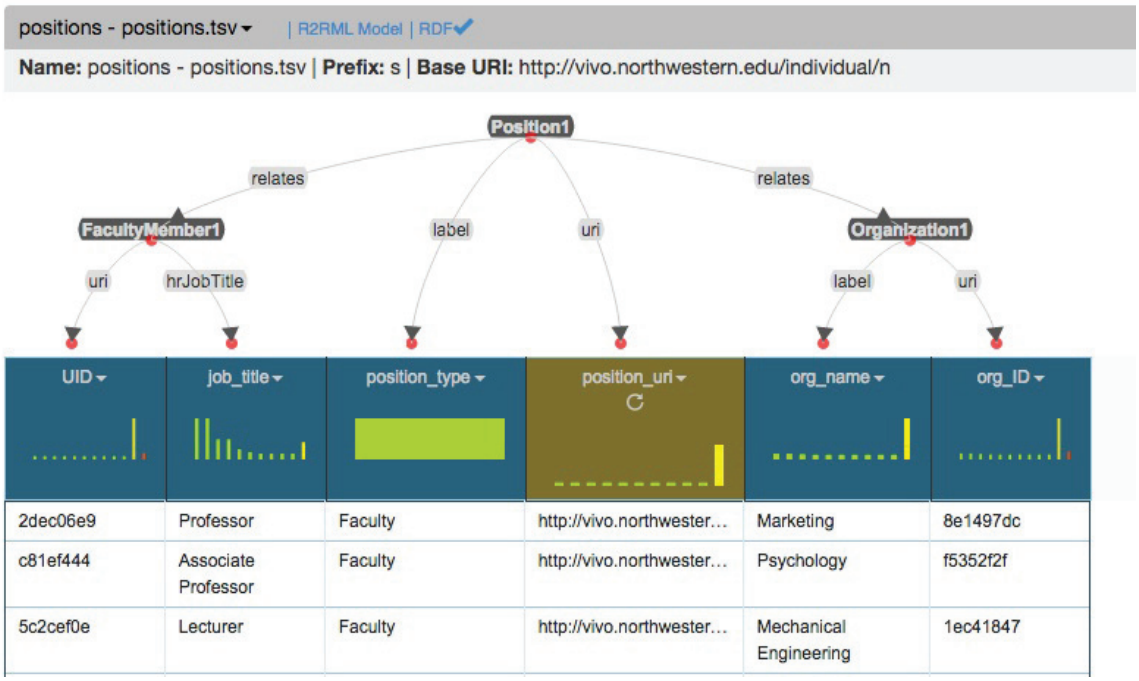


FIGURE 3.4 – Interface graphe d'édition des règles dans l'outil Karma [71].

Karma ne propose pas encore de modèle de règles intégrant les spécificités du domaine bibliographique comme les relations bibliographiques avancées. X3ML peut également être couplé à une interface appelée *3MEditor* qui permet, comme Karma, d'évaluer et de modifier les correspondances de transformation de manière visuelle. Dans le cas de la FRBRisation, la spécificité des catalogues bibliographiques implique systématiquement des échanges entre les experts du domaine et les experts informatiques. Pour cela, Karma et X3ML sont des solutions intéressantes pour la communauté. Cependant, les deux outils sont conçus de manière générique (ils ne proposent pas de modèle de règle prédéfini au contexte bibliographique comme XC, VFRBR ou FRBR-ML) et ont été essentiellement expérimentés sur le domaine des musées, où les œuvres ont moins d'interactions entre elles que dans l'univers bibliographique.

Améliorations particulières pour la FRBRisation

Parmi les outils présentés précédemment, certains proposent des fonctionnalités ou améliorations répondant à des enjeux spécifiques du domaine bibliographique. VTLS et XC implémentent toutes les deux des vocabulaires facilitant l'interopérabilité des données FRBR produites dans le Web Sémantique et la communauté documentaire. Ils utilisent par exemple le vocabulaire RDA

qui est une référence importante pour la diffusion de contenus bibliographiques dans le monde [32]. L'exploitation de champs non structurés dans les notices bibliographiques (*ex.*, MARC21 *added entries*) est une autre amélioration proposée par certains outils. Ces champs peuvent contenir des informations essentielles pour l'extraction des connaissances, mais la sémantique même du champ est abstraite et nécessite un algorithme plus avancé pour interpréter les données. XC et VFRBR peuvent par exemple analyser certains champs du format MARC21 pour créer de nouvelles entités FRBR (nouveaux contributeurs d'une Œuvre) si un certain motif d'information est présent. Ces deux outils proposent également des contributions intéressantes pour la modélisation des connaissances avec FRBR. XC propose le modèle *XC Schema* comme une augmentation du modèle conceptuel de FRBR en intégrant des spécificités particulières de MARC (*ex.*, MARC21 Holding records). VFRBR propose également une modélisation étendue de FRBR avec de nouvelles propriétés adaptées au contexte de la musique.

L'enrichissement des métadonnées avec des référentiels externes est une autre amélioration pour la FRBRisation. FRBR-ML implémente un module d'enrichissement sémantique avec des bases de connaissances du Linked Open Data [4] (LOD). En détectant la présence de certaines entités locales dans des bases externes, l'outil peut agréger de nouvelles connaissances aux entités FRBR comme une meilleurs précision sémantique d'une relation entre un Agent et une Œuvre (*ex.*, un contributeur est en fait un réalisateur). Cependant, ce processus soulève plusieurs défis et nécessite des fonctionnalités spécifiques comme la fusion d'entités dont FRBR-ML et XC implémentent un algorithme dédié.

Le projet de FRBRisation TelPlus, n'ayant pas livré d'outil à disposition de la communauté, a néanmoins proposé certaines améliorations importantes au processus de FRBRisation. Ils ont d'abord proposé un ensemble de métriques d'évaluation d'un catalogue bibliographique afin de prédire la qualité des données FRBR produites. Nous présentons ces métriques dans le prochain chapitre. Le projet TelPlus a également étudié une méthode pour améliorer les performances de la phase de déduplication (présentée ci-après) avec notamment une méthode de blocking pour réduire la quantité d'entités à comparer.

Classification des solutions de FRBRisation

Afin de livrer une vision plus globale des solutions présentées dans cette section, nous classons l'ensemble des solutions évoquées selon trois critères, *la technique de migration*, *l'expressivité du modèle cible*, et *les améliorations spécifiques*. La **technique de migration** correspond à la technique de FRBRisation de l'outil c'est à dire le regroupement de notices ou l'extraction des connaissances avec des règles. L'**expressivité du modèle** cible correspond à la capacité d'un outil à tenir compte des spécificités et de la richesse du modèle cible théoriquement visé. Les **améliorations spécifiques** concernent toutes les améliorations, en termes de processus ou d'algorithmes, qui permettent à un outil de répondre à des contraintes spécifiques de qualité ou de performance pour un projet de migration. Nous estimons que ces critères peuvent être suffisamment discriminants pour un projet de FRBRisation. Aussi, cette classification permet de facilement repérer les outils les mieux adaptés à un projet de FRBRisation. La figure 3.5 présente un exemple de cette classification, appliquée aux outils de FRBRisation présentés précédemment.

Dans la figure 3.5, nous classifions 11 outils de FRBRisation dans deux graphes où les racines représentent la technique de FRBRisation. Les solutions sont réparties selon l'expressivité du modèle FRBR dans la mesure où nous avons suffisamment d'informations pour en juger. Les solutions sous la bulle "*Sans règles prédéfinies ou connues*" ne peuvent pas être classées selon l'expressivité du modèle car elles ne proposent pas de règles initiales que nous pouvons évaluer ou

4. `https://fr.wikipedia.org/wiki/Linked_open_data`

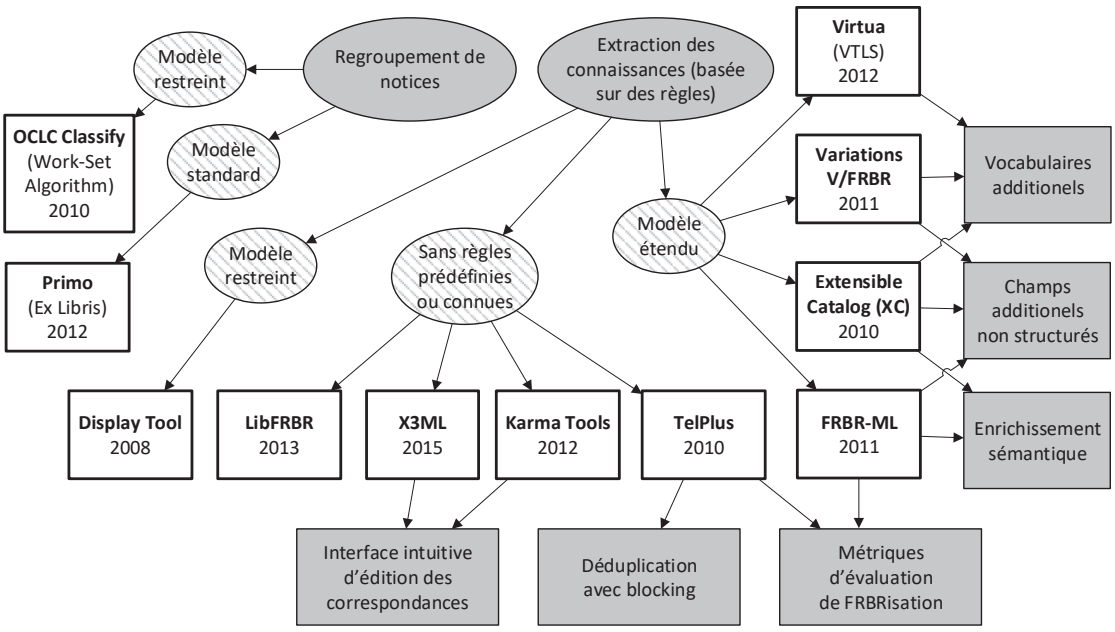


FIGURE 3.5 – Classification des solutions de FRBRisation où les ellipses grises symbolisent la technique de FRBRisation, les éllipses rayés représentent l'expressivité du modèle cible, les rectangles gris sont des améliorations spécifiques et les rectangles blanc sont des outils.

n'ont pas été expérimentées dans un contexte de FRBRisation. Les outils pointent vers certaines améliorations spécifiques que nous avons sélectionné pour leur pertinence au vue du contexte bibliographique. Avec cette représentation, il est donc possible d'identifier les deux outils qui intègrent un modèle d'enrichissement sémantique qui sont XC et FRBR-ML. Si, par exemple, nous sommes plutôt intéressés par des solutions d'extraction des connaissances qui ingèrent une interface de gestion des règles, notre classification nous oriente vers les solutions X3ML et Karma Tools. Il est important de remarquer que nous n'avons pas identifié de solution d'extraction des connaissances proposant ce type d'interface et intégrant un modèle adapté au contexte FRBR.

Dans cette section, nous avons classé les approches et outils de transformation des catalogues bibliographiques vers des bases de connaissances sémantiques reposant sur les principes FRBR. Nous avons observé un élan de propositions et d'amélioration des outils d'extraction des connaissances basées sur des règles. Cette technique, offrant d'intéressantes perspectives pour les institutions documentaires est également dépendante d'un processus additionnel de déduplication des entités produites pendant la transformation. Dans la prochaine section, nous étudions les solutions de déduplication qui sont appliquées ou extensibles au contexte bibliographique.

3.3 Déduplication

Nous étudions une autre étape du processus d'enrichissement de métadonnées bibliographiques appelée déduplication (*cf.*, figure 3.1). La déduplication correspond à la découverte automatique d'équivalences entre des entités du monde réel, afin de ne conserver et représenter que des entités uniques [28]. Dans un contexte de migration et d'enrichissement de métadonnées bibliographiques, cette phase permet de détecter et de fusionner les entités en doublons si plusieurs catalogues, incluant de la redondance, sont migrés ou si plusieurs sources de données, ayant des entités en commun, ont été utilisées pour réaliser l'enrichissement. Dans un processus de FRBRi-

sation basé sur la technique d'extraction des connaissances, qui traite chaque notice séparément, la déduplication assure la cohérence et la qualité de la base de connaissances produite [1].

3.3.1 Prérequis pour la déduplication

La phase de déduplication fait appel à deux processus très étudiés dans le domaine de l'intégration de données qui sont la détection d'entités équivalentes (*Entity Matching*) et la fusion de données (*Data Fusion*). Le processus d'Entity Matching calcule un niveau de similarité entre différentes entités en utilisant une combinaison de mesures de similarités (*cf.*, [29]) qui évaluent la proximité syntaxique et/ou sémantique des informations de ces entités [66]. Les particularités de la tâche d'Entity Matching sont largement détaillées dans les études de Elmagarmid *et al.*, [42] ou encore de Naumann et Herschel [93]. Le processus de fusion définit la manière dont les données de deux entités (jugées équivalentes selon le processus d'*Entity Matching*) doivent être conservées et/ou combinées [19]. Plus particulièrement, la fusion doit gérer le cas des données conflictuelles, c'est à dire lorsque les deux entités utilisent les même propriétés. Afin d'éviter l'ajout de redondance dans les données de l'entité finale, la fusion se déroule en deux temps. Dans le premier temps, le processus décide, en cas de propriétés conflictuelles s'il est nécessaire de résoudre le conflit, si les données peuvent cohabiter ou si les données ne sont pas considérées. Dans un deuxième temps, si un conflit doit être résolu, le processus applique une combinaison de méthodes permettant de décider quelle valeur sera retenue pour la propriété conflictuelle. La méthode de décision peut se baser sur des critères quantitatifs (*ex.*, la donnée apparaissant le plus souvent) mais également qualitatifs (*ex.*, une donnée jugée "vraie") [38].

La déduplication appliquée à de grands volumes de données peut poser des problèmes de performances car de nombreux calculs de similarité sont nécessaires. De plus, un projet de FRBRisation peut nécessiter plusieurs itérations pour la transformation des notices, impliquant à chaque fois de rejouer un processus coûteux de déduplication. C'est pourquoi, ce dernière intègre bien souvent une étape préliminaire de blocking, permettant de réduire le nombre de comparaisons d'entités à effectuer [39]. Les techniques de blocking ont été répertoriées dans cette large étude de Christen [27]. Le fonctionnement du blocking consiste à répartir les entités à comparer dans des "blocks" en considérant que deux entités issues de blocks différents ne doivent pas être comparées. La construction des blocks n'a donc pas seulement un impact sur les performances de la déduplication mais aussi sur la qualité du processus global. En effet, une composition trop restrictive des blocks peut résulter en de nombreux faux positifs.

Pour résumer la déduplication selon les trois processus présentés succinctement jusqu'ici, la figure 3.6 montre un exemple théorique de déduplication en illustrant les étapes principales. Dans cet exemple, la phase de *Blocking* génère des paires pertinentes d'entités à comparer. Dans l'exemple de la figure 3.6, le produit cartésien des entités à comparer, issues de bases différentes, donne quatre comparaisons ({A,C},{A,D},{B,C},{B,D}). Considérons par exemple que les entités A et D aient des attributs (de blocking) discriminants, la phase de blocking va supprimer cette potentielle paire de comparaison au processus de Matching. Cette seconde phase consiste ensuite à appliquer les mesures de similarités aux entités de chaque paire afin de définir si ces dernières sont équivalentes ou non. Pour illustrer cette phase, nous considérons par exemple que l'entité A de la base m est équivalente à l'entité C de la base n. Enfin, la phase de *Fusion* détermine, dans un cas d'équivalence, la manière dont les propriétés des entités à fusionner seront organisées. Ici, les propriétés k et l de C ne sont pas conservées au profit des propriétés x et y de A.

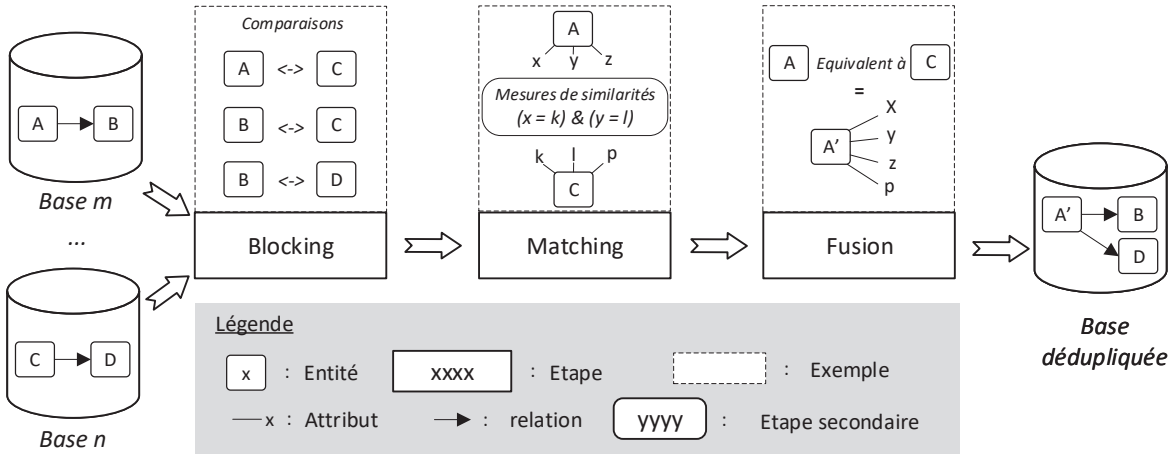


FIGURE 3.6 – Exemple schématique pour l'étape de déduplication

3.3.2 Déduplication d'entités bibliographiques

Il existe différentes approches d'*Entity Matching* pour la déduplication qui ont été analysées et comparées dans l'étude de Kopcke *et al.* [75]. Dans la majorité des cas, les solutions sont évaluées selon au moins un critère de qualité, généralement *F-Mesure* qui combine la précision et le rappel sur les données produites et comparées à une collection experte, et un critère de performance comme le temps de traitement. Dans le contexte des données bibliographiques, la pertinence de l'évaluation de tels outils est également influencée par les jeux de données utilisés. En effet, les expérimentations disponibles de déduplication de métadonnées documentaires se limitent principalement aux ressources scientifiques (*ex.*, articles ou journaux), [55, 67], car les jeux de données sont plus faciles d'accès (*ex.*, ACM-DBLP [5]). Or, les métadonnées scientifiques ne sont pas représentatives des données culturelles qui peuvent avoir beaucoup plus de corrélations entre-elles. C'est le cas par exemple dans le domaine de la lecture publique où de nombreux titres et noms de familles des auteurs peuvent être similaires. Dans ce cas, il est nécessaire de comparer plusieurs attributs des entités pour obtenir une meilleurs qualité de déduplication [84, 91].

La nécessité de combiner plusieurs attributs pour garantir une qualité suffisante dans la mesure de similarité des entités se retrouve dans les travaux de l'OCLC qui montrent que la déduplication des Œuvres FRBR doit être réalisée par la création de clés issues de plusieurs attributs comme différents titres, noms d'auteurs ou de contributeurs [61]. Dans l'étude de Kopcke *et al.* sur les approches d'Entity Matching, les solutions obtiennent de meilleurs résultats en termes de qualité (F-Mesure proche de 100%) lorsque plusieurs attributs sont utilisés pour le calcul de similarité sur des jeux de données bibliographiques [75].

Cependant, les performances des outils peuvent être fortement réduites lorsqu'un nombre plus élevé d'attributs et/ou plus de mesures de similarité sont impliqués dans le processus. Whang et Garcia-Molina proposent, dans le cas où plusieurs bases de données sources (*ex.*, plusieurs catalogues bibliographiques) sont concernées par la déduplication de bénéficier des alignements entre certaines sources pour accélérer le calcul des alignements restants [143]. Par exemple, si deux entités symbolisant des ouvrages, présentes dans deux catalogues bibliographiques différents, sont jugées équivalentes, alors on peut immédiatement considérer les deux entités *auteurs* des ouvrages comme équivalentes (c-à-d., c'est aussi le même auteur). Hogan *et al.*, proposent de bénéficier de relations d'équivalences existantes grâce aux ontologies des sources à dédupliquer pour inférer de nouvelles équivalences comme utiliser la propriété *sameAs* entre des sources A et

5. `https://www.openicpsr.org/openicpsr/project/100843/version/V2/view`

B et B et C afin d'inférer des équivalences entre A et C [65]. Dans le contexte de FRBRisation d'un catalogue isolé et spécialisé (mais d'un volume important), ces approches auront un impact limité sur les performances du processus. Il est, dans notre contexte, nécessaire de pouvoir aussi améliorer les performances de la déduplication sans recours à des bases de données externes.

Dans le domaine bibliographique, le blocking a été étudié pour la FRBRisation du projet TelPlus [82]. L'approche proposée consiste à définir un ensemble de clés pour chaque entité d'une classe de FRBR. Par exemple, une entité de classe Œuvre se voit attribuer plusieurs clés composées par exemple des titres, des auteurs ou des relations à d'autres entités. Chaque entité avec ses clés est considérée comme un bloc. Ensuite, l'approche fusionne les blocs qui partagent au moins une clé équivalente selon une fonction de hachage. Cette méthode de blocking ne considère pas les clés du WorkSet Algorithm qui permettrait de détecter des équivalences directement dans les blocs. De plus, certaines clés peuvent produire des blocs déséquilibrés avec de gros blocs limitant les performances de la phase d'Entity Matching (*ex.*, si un nom d'auteur est couramment utilisé).

Wang *et al.* proposent d'améliorer les performances du blocking en le réalisant de manière itérative [144]. En effet, plutôt que de traiter l'ensemble des blocs simultanément, les auteurs proposent de ne traiter qu'une partie des blocs à l'itération n et, après la fusion des entités équivalentes de cette itération, produire de nouvelles clés de blocking plus riches de cette fusion pour l'itération $n + 1$. D'autres approches utilisent l'apprentissage automatique et supervisé pour estimer la méthode optimale de blocking comme Bilenko *et al.*, [18] et plus récemment Papadakis *et al.*, [101]. Ces approches définissent des traits (appelés *features* en anglais) et fournissent un jeu de données expert permettant de réaliser un apprentissage des meilleurs combinaisons de traits selon les données. L'inconvénient des approches itératives ou par apprentissage réside dans les efforts d'implémentation ou de paramétrage qu'il est nécessaire de réaliser en amont de la déduplication.

Une autre méthode permettant d'augmenter les performances du blocking consiste à forcer l'homogénéisation de la taille des blocs afin de maximiser le calcul parallèle de ces blocs, dans le cas où ces derniers sont déséquilibrés. Kolb *et al*, proposent la technique de blocking appelée *PairRange* qui consiste à construire des plages équilibrées de calculs parallèles de paires d'entités à comparer, à partir des blocs non-équilibrés obtenus par une méthode plus classique de blocking [74]. La taille des plages est définie arbitrairement, bien souvent en fonction des possibilités en termes de calcul parallèle de la ou les machine(s) utilisée(s). Ensuite, l'algorithme attribue un index à chaque paire à la suite d'un processus d'énumération des entités d'un bloc non-équilibré. Cet index permet d'affecter les paires à une plage de calcul tout en respectant l'équilibre des charges. Pour cela, *PairRange* est basé sur le paradigme MapReduce où deux premières tâches Map puis Reduce servent à construire une matrice des blocs initiaux (non équilibrés) de paires à énumérer, puis deux autres tâches Map et Reduce servent respectivement à affecter les entités dans des blocs virtuels (Map) puis, à partir de ces nouveaux blocs, à reconstituer les pairs d'entités à comparer dans une plage de calcul (Reduce). La figure 3.7 montre un exemple d'énumération de paires pour des entités nommées *A, B, ..., O*.

Sur la Figure 3.7, on observe quatre blocs créés à partir de quatre clés de blocking w, y, x et z. Dans ces quatre blocs, l'algorithme d'énumération attribue un chiffre à chaque paire d'entités dans les blocs (*ex.*, 3 pour HB, 11 pour MF). Considérant une parallélisation en 3 tâches de calcul, (représentée par les nuances de gris), *PairRange* crée trois plages de calcul de taille 7, 7 et 6 respectivement de 0 à 6, de 7 à 13 et de 14 à 19. Le chiffre obtenu par énumération des paires, c'est à dire l'index de paire, définit la plage dans laquelle la paire doit être affectée pour être évaluée par une tâche de calcul parallèle. Au final, au lieu d'utiliser les blocs initiaux déséquilibrés (*ex.*, bloc Φ_1 de taille 1 et bloc Φ_3 de taille 10), le calcul s'effectue sur 3 pages plus équilibrées. La méthode PairRange est adaptée aux processus de déduplication nécessitant

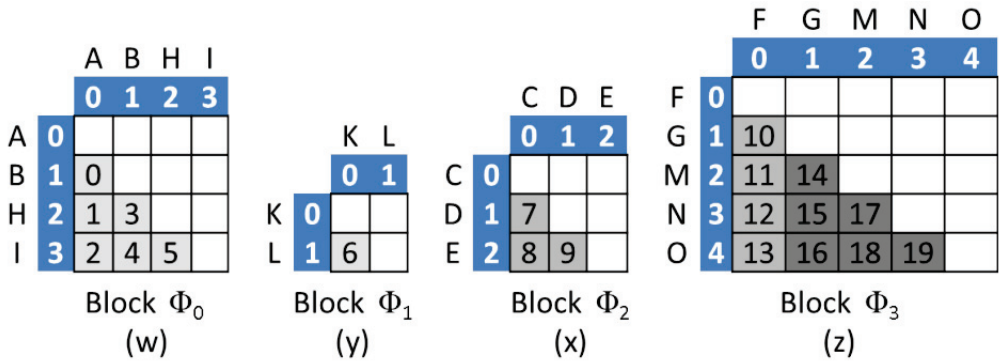


FIGURE 3.7 – Exemple d'énumération des pairs dans PairRange, extrait de [74]

d'importants calculs pour la comparaison des entités car la réorganisation des blocs initiaux en blocs équilibrés peut avoir un coût non négligeable si les données sont très hétérogènes.

Dans le contexte de la FRBRisation, la déduplication est nécessaire quand les URIs des entités produites par transformation des notices ne reposent pas sur des identifiants permettant une comparaison simple des entités (*ex.*, ISBN [6], titre uniforme, ARK [7]). Différentes méthodes et principes ont été proposé pour réaliser une comparaison des entités FRBR avec une bonne qualité (*ex.*, clés du WorkSet Algorithm). Le principal problème réside dans les performances de la déduplication quand celle-ci doit être réalisée de multiples fois (c'est souvent le cas). Cette performance est fortement dépendante soit des critères de blocking s'ils sont définis à l'avance par les experts, soit des traits d'apprentissage si les critères sont appris automatiquement. Dans tous les cas, une réflexion sur ces critères et ces traits est une étape systématique et nécessaire lors d'un processus de FRBRisation d'un catalogue comportant de nombreuses notices, ce qui n'est pas en phase avec le besoin d'industrialisation de la migration des notices dans la communauté.

3.4 Enrichissement sémantique d'entités FRBR

Nous avons vu dans la section 3.2 que certains outils de FRBRisation proposent un module d'enrichissement sémantique. Cette fonctionnalité présente un avantage important pour préserver, voir améliorer, la qualité du catalogue transformé. Cependant, cette tâche soulève de nombreux défis issues du domaine de l'intégration de données. Dans cette section, nous étudions les approches d'enrichissement sémantique appliquées aux métadonnées bibliographiques FRBR. La FRBRisation étant un processus de transformation de métadonnées, l'enrichissement sémantique, dans notre contexte, consiste à intégrer de nouvelles métadonnées aux entités FRBR. Cette agrégation de nouvelles connaissances implique l'interrogation de sources externes contenant potentiellement des informations sur les entités à enrichir. C'est pourquoi, la première étape du processus consiste à aligner les entités locales avec les entités distantes. Cette tâche est connue sous le nom d'Entity Linking [117]. Différentes approches sont détaillées dans les études de Dai *et al.* [34].

La méthode d'enrichissement proposée par l'outil FRBR-ML interroge des sources structurées du web de données (*ex.*, DBPedia) pour lier les entités FRBR locales [128]. L'alignement avec les entités distantes est réalisé de différentes manières comme la reproduction des URIs distantes à partir des données locales, l'interrogation de la source distante via un *endpoint* SPARQL [8]

6. `https://fr.wikipedia.org/wiki/International_Standard_Book_Number`
7. `http://www.bnf.fr/fr/professionnels/issn_isbn_autres_numeros/a.ark.html`
8. `https://www.w3.org/wiki/SparqlEndpoints`

ou encore l'utilisation d'APIs de recherche. La création de requêtes et d'URIs pour l'enrichissement peut mener à de nombreuses erreurs de désambiguïsation. De Wilde et Hengshen proposent d'améliorer le processus d'Entity Linking en tenant compte des différentes langues dans lesquels les données locales peuvent être stockées afin d'adapter les requêtes envoyées aux différentes sources de données distantes [145]. Frontini *et al.* traitent le problème de désambiguïsation des personnes en proposant une méthode pour mieux distinguer certaines entités ayant de nombreux homonymes (*ex.*, Victor Hugo) [50]. La combinaison de ces solutions, ou l'utilisation d'outils d'alignement spécialisés comme DBPedia Spotlight (*cf.*, [87]), permet d'obtenir de bons résultats pour l'alignement des entités locales avec le web de données.

Les approches existantes d'Entity Linking reposent essentiellement sur un alignement avec des données distantes structurées en RDF [109]. Cependant, dans le contexte documentaire, le principal problème vient de la faible disponibilité de données bibliographiques dans une forme structurée (*ex.*, en RDF) et du fait qu'aucune solution libre d'enrichissement de données bibliographiques n'exploite de sources non structurées. Il y a encore un grand nombre d'informations sur le web qui nécessitent d'être préalablement structurées, par des solutions de type DeepDive [118, 99, 92], avant d'être exploitées. Toutefois, la standardisation des identifiants ISNI pour les entités personnes, l'utilisation plus importante des clés ARK (*cf.*, [10]) pour les entités FRBR ou encore les APIs de bibliothèques qui fournissent des données en RDF (*ex.*, en France [134]) vont progressivement faciliter l'étape d'Entity Linking dans les futurs processus de migration de données.

Lorsque des entités FRBR sont alignées avec des sources externes, il est possible d'enrichir les informations de ces entités. Parmi les solutions existantes d'extraction de connaissances, les approches de Haslhofer *et al.*, [57] ou Chianese *et al.*, [26] réalisent l'intégration des nouvelles données selon le paradigme du monde "ouvert", c'est à dire que toutes les informations disponibles sont intégrées. Dis autrement, on ne tient pas compte de la nature ni de la structure des données distantes. Dans le cas de modèles utilisant RDF, les triplets distants sont directement ajoutés aux triplets locaux. Dans l'approche de Haslhofer *et al.*, les utilisateurs peuvent analyser et annoter les résultats *a posteriori* via une interface dédiée. Ce type d'extraction, aussi appelée *Open Information Extraction* (*cf.*, [44, 41]) peut être difficile à appliquer dans le monde bibliographique car les métadonnées restituées aux utilisateurs doivent respecter une structure particulière afin de ne pas créer de confusion à la lecture de ces informations (*ex.*, respect des niveaux sémantiques de FRBR). Toutefois l'aspect communautaire peut avoir un intérêt pour certaines institutions. Dans d'autres travaux comme [135] ou [126], l'extraction des connaissances est réalisée en monde "clos", c'est à dire que le périmètre et la structure des données à extraire sont préalablement définis dans un modèle d'enrichissement. Ce modèle doit contenir un ensemble de règles permettant de construire des requêtes spécifiques afin de n'intégrer que certaines connaissances précises.

L'inconvénient de l'enrichissement en monde "clos" est qu'il implique une connaissance préalable des modèles de toutes les sources distantes, afin de pouvoir construire correctement les requêtes. Une solution à ce problème consiste à réaliser un processus préalable d'alignement de modèle (Ontology Matching) entre le modèle local et les modèles distants, si la description des ontologies distantes est disponible [15]. Si les modèles distants ne sont pas explicités, le processus d'Ontology Matching peut être réalisé en comparant directement les instances des bases locales et distantes pour en déduire les similitudes des modèles (Instance-based Matching) [141]. Des outils comme *LogMap* [68] ou *AgreementMakerLight* [46] implémentent différentes méthodes d'alignement de modèles pour obtenir un niveau de qualité satisfaisant. Toutefois, dans le domaine documentaire, les processus d'Ontology Matching peuvent être difficiles à réaliser car les modèles bibliographiques qui émergent dans le web de données sont encore récents ou, pour certains, très complexes à aligner (*ex.*, dans le domaine de la musique avec FRBRoo [113]). Une méthode alternative a été proposée par Nuzzolese *et al.*, qui consiste à analyser les données ex-

traites de sources distantes, selon la méthode d'*Open Information Extraction*, puis de réaliser des statistiques sur la fréquence d'utilisation de certaines propriétés en fonction du type de ressource décrite afin de déduire des motifs de description des données, appelés *Encyclopedic Knowledge Patterns* [96]. Cette méthode facilite la découverte des modèles de données distants et réduit les efforts nécessaires à la modélisation des connaissances issues des sources externes.

Nous avons identifié d'autres approches d'enrichissement des métadonnées qui peuvent être pertinentes dans le contexte des données documentaires. Les travaux de Lacasta *et al.*, permettent d'améliorer les résultats de recherche dans un catalogue en proposant des alignements plus riches entre les concepts, issus de thésaurus et associés aux informations bibliographiques (*ex.*, auteurs, sujets, genres), et les ontologies bibliographiques [78]. En alignant les définitions des concepts de thésaurus avec un dictionnaire comme WordNet puis à une ontologie de haut niveau comme DOLCE (ou FRBR), l'approche permet de mieux interpréter les termes saisis par les utilisateurs. Suivant le même objectif, les travaux de Hinze *et al.*, permettent également un enrichissement des concepts associés aux notices bibliographiques afin de laisser l'utilisateur, quand c'est possible, préciser la sémantique exacte des termes de recherche qu'il a saisit [62, 63]. Avec ces méthodes, l'utilisateur est plus impliqué dans son processus de recherche, dont le moteur devient lui-même plus interactif. West *et al.*, étudient une manière différente d'enrichir les données en transformant les métadonnées locales en questions formulées en langage naturel qui sont soumises à des banques de questions/réponses sur des sources du web de données (Wikidata, DBPedia) [142]. L'intérêt de l'approche, est qu'ils utilisent les données locales ainsi que des sources externes pour enrichir les termes inclus dans les questions afin de mieux cibler la réponse la plus pertinente quand elle existe. Ces différentes approches reposent toutes sur un principe d'enrichissement des termes qui sont associés aux métadonnées bibliographiques. Cette méthode a de nombreux avantages pour offrir de meilleurs fonctionnalités de recherche à l'utilisateur ou pour créer des rapprochements intéressants entre des ressources documentaires.

3.5 Discussion et conclusion

Notre étude des techniques et solutions pour la transformation des catalogues bibliographiques vers FRBR ainsi que sur l'enrichissement des entités produites nous montre qu'il existe de multiples approches et outils offrant diverses fonctionnalités pour réaliser une transformation des anciens catalogues de notices en respectant les enjeux de qualité inhérents au domaine documentaire. Les possibilités de personnaliser les vocabulaires du modèles cible, de réaliser une phase additionnelle d'enrichissement sémantique ou de bénéficier d'une interface intuitive d'édition des règles de transformation et d'enrichissement sont des améliorations très utiles pour faciliter la réalisation de projets d'adoption des principes FRBR dans les institutions documentaires. Nous observons cependant qu'il n'existe pas encore d'outil spécialisé dans le domaine documentaire qui intègre toutes ces fonctions. Toutefois, nous estimons qu'il sera rapidement possible de combiner les améliorations existantes dans de futurs outils spécialisés. Les questionnements que nous soulevons alors résident dans l'exhaustivité des fonctionnalités que nous avons observées pour réaliser une solution idéale de migration et d'enrichissement des catalogues documentaires. Nous rappelons que dans le contexte de la thèse, les projets de FRBRisation sont encore non envisageables par une grande majorité des institutions faute de moyens techniques et d'expertise sur le sujet. C'est pourquoi, nous cherchons à identifier les améliorations qui sont réellement nécessaires pour faciliter la mise en place de ce type de projets dans ces institutions.

Des réponses à ces questionnements viennent non pas des outils eux-mêmes mais des expérimentations, au sens large, d'adoption de modèles sémantiques dans des domaines où les données sont issues de formats difficilement exploitables. Nous observons que dans ces expérimentations, la

modélisation des connaissances est un enjeu majeur pour la réussite du projet car elle impacte la manière dont les données vont être restituées aux utilisateurs mais aussi la façon dont ces données vont être intégrées et enrichies [90]. Les expérimentations de Szekely *et al.* avec la solution Karma ont montré la nécessité de développer des outils complémentaires (aux outils de migration des notices) pour assister les experts de la migration dans la modélisation préalable des connaissances et des vocabulaires associés [125]. D'autres travaux comme [97], [135] ou encore [7] ont eux recours à la création de méta-modèles pour faciliter l'intégration de nouvelles connaissances dans des systèmes d'informations culturels. Dans ce même objectif, *Aalberg* propose d'encapsuler les règles de modélisation de connaissances des musées dans des "méta-motif" [4].

Le contexte d'évolution des normes et pratiques dans les communautés culturelles (*cf.*, chapitre 2) accentue également le recours à ces travaux de création de nouveaux modèles de données en amont des projets de transformation des métadonnées. En conséquence, nous observons un réel besoin de méthodologie et de spécifications dans le contexte de la modélisation des connaissances afin de mieux impliquer les experts du domaine documentaire dans la transition des institutions culturelles vers le Web Sémantique [102, 8]. Par ailleurs, certaines phases majeures des processus de migration et d'enrichissement sont encore réalisées de manières isolées ou spécifiques comme comme la phase de *blocking*, la création des règles de migration ou encore la validation des entités produites. Ces étapes peuvent être partiellement industrialisées en formalisant les standards du domaine bibliographique qui se mettent progressivement en place. L'objectif est de réutiliser l'existant en matière de modélisation bibliographique (*cf.*, [133, 79]) ainsi que les métriques proposées par la communauté (*cf.*, [82, 127]) pour mieux assister les experts documentaires dans la préparation d'un projet de transformation des métadonnées. Sur la figure 3.8 nous synthétisons ces perspectives d'amélioration en annotant le schéma du processus initial de transformation et d'enrichissement des notices avec des processus et interactions supplémentaires (détails ci-après).

Dans ce schéma, nous proposons d'intégrer un ensemble de caractéristiques des données bibliographiques (en bas à gauche) qui peuvent faciliter la réalisation de certaines étapes clés du processus global. Ces caractéristiques sont composées de métriques d'évaluation permettant l'analyse des spécificités d'un catalogue bibliographique et d'une base de connaissances FRBR ainsi que des motifs de modélisation issus des standards du domaine (*ex.*, la modélisation d'une traduction d'Œuvre respecte toujours le même motif FRBR). Grâce à ces informations, l'expert peut bénéficier de l'analyse automatique du catalogue et des motifs prédéfinis pour réaliser une modélisation conceptuelle (que nous appelons aussi *méta-modèle*) de sa future base de connaissances bibliographiques (sans nécessiter de compétences informatiques). L'étape appelée *Instanciation* sur la figure 3.8 consiste ensuite à traduire cette modélisation conceptuelle en un modèle pratique de migration et d'enrichissement des notices du catalogue. Plus concrètement, il s'agit d'ajouter les bonnes conditions et fonctions nécessaires à l'interprétation des notices et à la bonne intégration des données dans la future base de connaissances. Une fois encore, l'analyse préliminaire du catalogue permet d'extraire les spécificités des données issues des notices afin de faciliter cette étape d'ajout de règles et de conditions à la modélisation conceptuelle. Les règles de migration et d'enrichissement, une fois définies, sont utilisées par les deux processus de migration et d'enrichissement. Ces derniers ayant tous les deux recours à un processus de déduplication, nous proposons également de simplifier le travail de l'expert en automatisant le paramétrage de la phase de blocking, en amont de la déduplication, afin de réduire le temps passé par l'expert sur des considérations essentiellement informatiques.

Il est important de noter que cette vision originale que nous apportons au processus global de migration et d'enrichissement de notices bibliographiques s'inscrit comme un complément (et non une rupture) des propositions et outils existants dans la communauté. La méta-modélisation des connaissances standards, en amont de l'écriture des règles, doit simplifier le travail du documen-

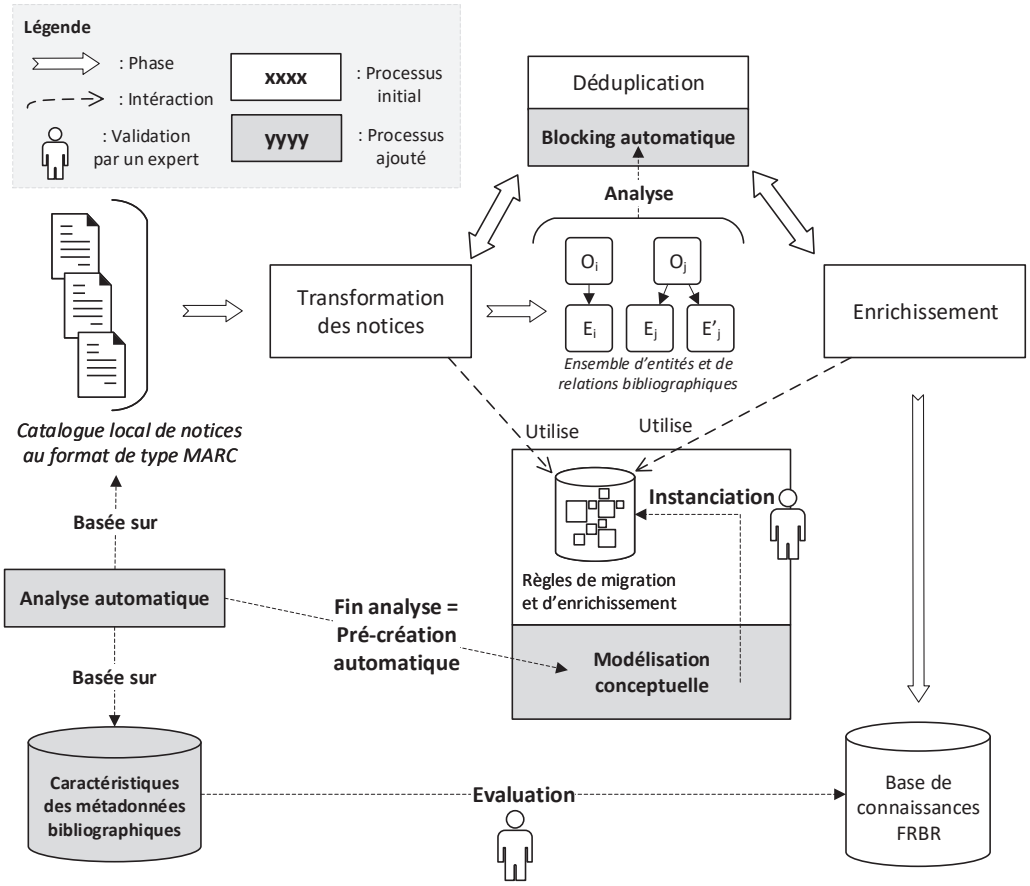


FIGURE 3.8 – Version étendue du processus théorique de transformation et d'enrichissement de notices bibliographiques selon notre approche

taliste en augmentant le niveau d'abstraction des règles au niveau des motifs bibliographiques dont il maîtrise la terminologie. L'instanciation de cette modélisation pour réaliser la transformation peut ensuite être implémentée dans un outil existant si ce dernier fournit des mécanismes suffisamment avancés de modélisation (comme X3ML ou Karma Tools). Dans le reste de ce document, nous détaillons les contributions que nous apportons à ce domaine selon cette vision.

Chapitre 4

Évaluer la migration de notices bibliographiques

Ce chapitre décrit trois ensembles de métriques dédiées à l'évaluation d'un processus de migration de données bibliographiques et présente un benchmark original d'évaluation des solutions informatiques dans ce domaine. Les métriques proposées permettent notamment d'évaluer les phases d'interprétation et de transformation des données issues d'anciennes notices documentaires vers des bases de connaissances bibliographiques. Notre benchmark intègre ces métriques ainsi que des jeux de données dédiés à une évaluation caractéristique des solutions de migration.

Sommaire

4.1 Introduction

La communauté documentaire préconise l'adoption des principes du Web Sémantique ainsi que de nouvelles ontologies et vocabulaires pour la gestion des métadonnées bibliographiques [54]. La standardisation de modèles bibliographiques comme FRBR témoigne de cet élan [114, 119, 88]. Cette adoption repose cependant sur un processus complexe qui implique trois défis majeurs qui sont (1) l'interprétation des connaissances bibliographiques [5], qui est modélisée dans les notices

existantes, (2) la migration des nombreuses métadonnées vers de nouvelles bases de connaissances sémantiques [82] et (3) la valorisation et la réutilisation de ces nouvelles bases [90].

Face à ces différents enjeux, les méthodes et outils pour la transformation des métadonnées bibliographiques sont au cœur des considérations des professionnels du domaine documentaire [152]. Les solutions existantes ont pour objectif de simplifier le travail des experts en facilitant la création des règles de migration et en traitant automatiquement de grands volumes de notices [35]. Cependant, nous avons vu au chapitre 3 que les spécificités des métadonnées bibliographiques impliquent des processus spécifiques pour répondre à ces objectifs comme l'interprétation des motifs de connaissances, ou l'intégration de données d'enrichissement issues de sources externes. Dans le paysage prolifique des outils de migration de métadonnées, il est devenu nécessaire, pour les institutions documentaires, de pouvoir évaluer et comparer les solutions à leur disposition.

4.1.1 Contexte des travaux

Dans le contexte de cette thèse, nous avons observé les difficultés soulevées par la communauté documentaire pour évaluer les méthodes et outils dédiés à la migration des catalogues bibliographiques vers des bases de connaissances sémantiques. En effet, une majorité des outils ou approches ont été expérimentés dans le cadre de projets spécifiques et, en conséquence, les outils développés pour ces occasions demeurent rarement disponibles ou même réutilisables et les jeux de données utilisés (c-à-d., des sous-ensembles de notices) sont bien souvent indisponibles [105]. Dans d'autres cas, seuls quelques exemples de jeux de données d'un projet de migration sont fournis à titre d'illustration mais ne reflètent pas la complexité des catalogues bibliographiques [130]. Enfin, les métriques proposées pour évaluer ce type de projets de migration ne sont pas suffisantes pour anticiper toutes les spécificités des métadonnées bibliographiques [127, 82]. Nous avons donc orienté nos travaux pour considérer les critères d'évaluation qui permettent de mieux anticiper les problèmes de ce domaine et ainsi réduire les coûts des projets de migration.

Dans ce chapitre, nous présentons des métriques permettant d'évaluer un processus d'intégration de données bibliographiques selon trois aspects. Le premier aspect concerne l'*interprétation* des connaissances bibliographiques issues des données en entrée (*ex.*, notices). Ces métriques permettent notamment à un expert d'évaluer les efforts nécessaires à l'adaptation d'un modèle de règles existant pour l'interprétation d'un catalogue donné. Le deuxième aspect traite des performances d'un outil utilisé pour transformer des notices en une base de connaissances sémantiques. Cet aspect permet notamment de mesurer si un outil répond aux délais de traitement accordés à des phases d'intégration automatique de données. Le troisième aspect concerne la qualité de la base de connaissances produite à l'issue d'un processus d'intégration de données. Les métriques proposées pour cet aspect permettent notamment d'évaluer si les métadonnées bibliographiques, qui sont ajoutées dans la nouvelle base de connaissances, répondent aux exigences de qualité et de réutilisabilité des institutions concernées. Nous intégrons ces métriques dans un benchmark avec lequel nous fournissons des jeux de données experts, contenant des données réelles, permettant d'évaluer différents aspects clés des solutions existantes de FRBRisation. Nous présentons des expérimentations réalisées avec ce benchmark sur trois outils récents de notre état de l'art.

4.1.2 Catégories de métriques

Les métriques présentées dans ce chapitre couvrent l'ensemble du processus d'intégration de données pour former une base de connaissances bibliographiques [1]. Ce processus est décomposé en trois grandes phases que nous appelons (1) l'*interprétation* des connaissances, (2) la *transformation* automatique de métadonnées bibliographiques et (3) la *validation* de la qualité de la base de connaissances produite. Pendant la phase d'interprétation, les experts analysent les

métadonnées en entrée et modélisent la future base de connaissances. Cette phase peut intégrer des sous-étapes de nettoyage des données et de configuration des outils qui seront utilisés pendant le processus global d'intégration des données. Dans le cadre de la migration de notices bibliographiques, l'étape la plus cruciale consiste à rédiger et organiser les règles de migration. La phase suivante de transformation utilise des outils qui prennent en entrée les données à intégrer et produit une base de connaissances sémantiques. Plus particulièrement, ces outils appliquent l'ensemble des règles, rédigées à la phase précédente, sur les données pour produite des entités et propriétés. De plus, un processus de déduplication (*cf.*, Chapitre 3) est employé pour éviter la redondance dans les données. Enfin, la phase de validation consiste, en premier lieu, à appliquer différents sous-processus optionnels, tel que l'alignement des métadonnées avec des sources externes, puis à valider et corriger les entités et propriétés qui ont été intégrées.

Nous faisons remarquer que la plupart des approches existantes d'intégration de données bibliographiques n'évaluent que la dernière phase du processus, c'est à dire qu'ils évaluent les données produites une fois la transformation réalisée entièrement. Dans nos travaux, nous supposons que l'évaluation en amont de la première phase d'interprétation des connaissances peut avoir un impact très positif sur la qualité de la future base de connaissances, et cela peut également permettre d'éviter des itérations inutiles de transformation de données. Concernant cette phase de transformation, son évaluation étant essentiellement focalisée sur les performances de l'outil employé, son impact reste fortement lié au contexte du projet d'intégration en question.

Dans ce chapitre, la section 4.2 détaille les métriques dédiées à l'interprétation des connaissances, la section 4.3 présente les métriques pour la transformation automatique des métadonnées et la section 4.4 définit les métriques liées à la qualité de la nouvelle base de connaissances sémantiques. La section 4.5 présente notre benchmark permettant d'évaluer les solutions de FRBRisation et son application en considérant trois outils récents. La section 4.6 conclut ce chapitre.

4.2 Évaluer l'interprétation des connaissances bibliographiques

Pour rappel, le terme d'*interprétation*, dans notre contexte, correspond à une phase d'analyse des données permettant de déduire des règles pour extraire les informations utiles qui y sont représentées. Nous avons indiqué dans le chapitre 2 que les connaissances bibliographiques, issues des catalogues documentaires, représentent des motifs de relations entre entités qui sont implicitement décrits dans différents champs des notices. De plus, les champs utilisés pour décrire ces relations peuvent être gérés selon différentes pratiques de catalogage, qui ont pu évoluer au cours du temps. C'est pourquoi, l'interprétation de données bibliographiques, de manière automatique, peut entraîner de multiples erreurs et altérer la qualité des informations lors de la migration d'un catalogue. Pour mieux anticiper cette problématique, cette section décrit des métriques relatives aux spécificités des métadonnées bibliographiques. Les métriques proposées concernent, à la fois, l'analyse des pratiques de catalogage ainsi que la détection de relations bibliographiques avancées. De plus, ces métriques permettent d'analyser un jeu de règles, issu d'un outil de migration existant, afin d'évaluer si un catalogue peut être transformé correctement par cet outil.

Les métriques qui sont présentées ci-après retournent une valeur de pourcentage calculée à partir d'un nombre de notices, concernées par une spécificité ou une erreur, que nous divisons au nombre total de notices du catalogue à intégrer. Ainsi, si une métrique retourne une valeur proche de 100%, c'est que la grande majorité des notices sont concernées par la relation bibliographique ou l'erreur en question. Dans ce cas, la non-considération de cette spécificité ou erreur dans un jeu de règles d'intégration peut conduire à une mauvaise qualité des métadonnées intégrées.

4.2.1 Détection des relations bibliographiques implicites

Les relations bibliographiques, comme présentées au chapitre 2, sont des relations entre des entités bibliographiques qui représentent une information utile aux utilisateurs d'un catalogue documentaire. Par exemple, dans une migration de métadonnées vers FRBR, une préface ajoutée à un ouvrage doit introduire une nouvelle relation sémantique entre une entité de classe Œuvre (l'ouvrage) et une entité de classe Expression (la préface). Des détails sur ces relations sont fournis ci-après. Nous qualifions les relations bibliographiques d'implicites car, dans le contexte des notices bibliographiques, ces relations ne sont pas exprimées en termes de triplets explicites (sujet, prédicat, objet) mais en de multiples combinaisons de clés et valeurs (champs) qu'il convient d'interpréter. Dans la suite de cette partie, nous évaluons la détection de cinq catégories de relations que nous nommons respectivement **CORE** (relations élémentaires), **AUG** (augmentations), **DER** (dérivations), **AGG** (agrégations) et **COM** (relations complémentaires).

Définition 4.2.1 (CORE : Relations élémentaires)**.** . Les relations élémentaires sont nativement représentées dans chacune des structures d'informations (*ex.*, notices) à intégrer.

L'interprétation des relations élémentaires est propre à chaque projet d'intégration et ne peut pas faire l'objet d'une formalisation générale. Par exemple, dans le contexte d'une migration de notices bibliographiques vers FRBR, nous pouvons considérer que chaque notice à interpréter décrit au moins une Œuvre (la création originale), en relation avec au moins une Expression (la réalisation), elle-même en relation avec au moins une Manifestation (la publication). Si les notices décrivent des œuvres scientifiques ou artistiques, ces relations peuvent être très différentes (ces aspects sont détaillés dans le chapitre 5). C'est pourquoi, nous ne nous risquons pas à proposer une définition générale ici. Les quatre prochaines catégories représentent des motifs plus génériques qui peuvent être identiques entre différentes sources de données documentaires. Nous présentons donc, pour chaque catégorie, une définition textuelle, des illustrations et nous proposons un exemple de notation formelle pour la détection automatique des relations du motif en question. Nous commençons par étudier la détection de relations d'augmentation d'une Œuvre.

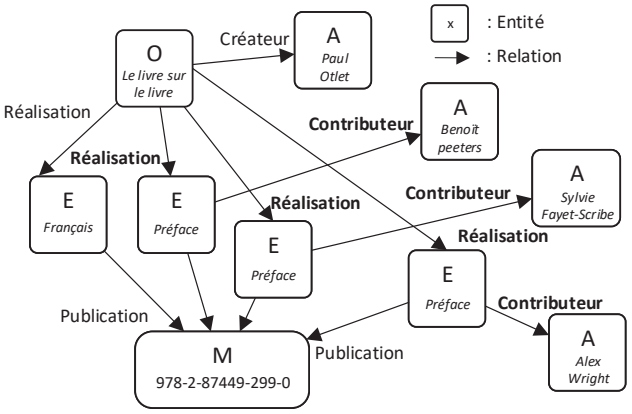
Définition 4.2.2 (AUG : Relations d'augmentation)**.** . La métrique AUG détecte le pourcentage de notices incluant une ou plusieurs relations d'augmentation, c'est à dire la description de contenus intellectuels supplémentaires qui ont été créés pour augmenter une Œuvre.

Les objets documentaires comme des illustrations ou des préfaces sont considérés comme des augmentations d'œuvres. Pour comprendre l'interprétation de ces dernières, la figure 4.1 présente un exemple de notice bibliographique (figure 4.1a) contenant des relations d'augmentation relatives à trois préfaces qui ont été ajoutées dans une édition du traité de documentation de *Paul Otlet*. De plus, nous proposons une modélisation de ces relations dans la figure 4.1b.

Dans cet exemple de la figure 4.1, nous observons que les préfaces peuvent être interprétées de différentes manières. D'abord, elles peuvent être déduites par le code de fonction (702$4 en UNIMARC) associé aux responsabilités secondaires qui, dans cet exemple avec le code *080*, signifie "*auteur d'une préface*". Ces augmentations peuvent également être interprétées par l'analyse d'autres données descriptives comme par exemple l'énumération des auteurs et contributeurs sur la couverture d'un livre (champ 200$g en UNIMARC). Ces informations peuvent indiquer, en langue naturelle, la présence d'une ou plusieurs préfaces. Les illustrations et autres types d'augmentations sont essentiellement déduites par le même procédé. La modélisation des augmentations se traduit par une relation sémantique entre l'Œuvre augmentée et son augmentation qui peut elle-même être une Œuvre (selon l'importance que l'institution lui accorde) ou une Expression. Dans l'exemple de la figure 4.1, les préfaces de l'Œuvre (*Le livre sur le livre*) sont modélisées comme des Expressions (avec leurs contributeurs respectifs) qui se matérialisent dans une seule et même Manifestation (le livre physique avec son identifiant ISBN).

010 $a 978-2-87449-299-0
101 $a fre
200 $a Le livre sur le livre
$b Texte imprimé $e traité de documentation
$f Paul Otlet
$g préfaces de Benoît Peeters,
Sylvie Fayet-Scribe et Alex Wright
540 $a Le traité de documentation : le livre sur le livre
606 $a Sciences de l'information
700 $a Otlet $b Paul $f 1868-1944 $4 070
702 $a Peeters $b Benoît $f 1956-.... $4 080
702 $a Fayet-Scribe $b Sylvie $f 1956-.... $4 080
702 $a Wright $b Alex $f 1966-.... $4 080

(a) Notice UNIMARC d'augmentation



(b) Exemple de modélisation FRBR

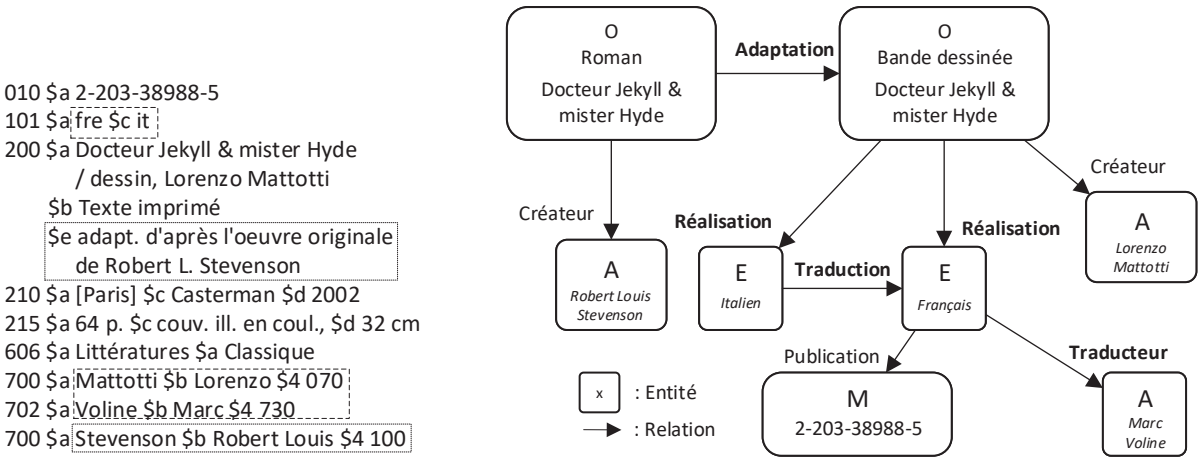
FIGURE 4.1 – Exemples de relations d'augmentation concernant des préfaces d'un ouvrage

Définition 4.2.3 (DER : Relation de dérivation)**.** . La métrique DER détecte le pourcentage de notices décrivant des relations de dérivation. Une dérivation concerne une Œuvre originale dont le contenu intellectuel a été modifié ou adapté pour produire une nouvelle Œuvre à part entière ou une nouvelle Expression de l'Œuvre originale.

L'interprétation des dérivations est similaire aux augmentations. Cependant, il peut exister des cas spécifiques aux dérivations qu'il convient de distinguer. La figure 4.2 illustre ces spécificités avec un nouvel exemple de notice (figure 4.2a) pour laquelle nous proposons une suggestion d'interprétation (figure 4.2b). Dans cet exemple, il s'agit de la description d'une bande dessinée issue d'une traduction (italien vers français) et qui est une adaptation du roman nommé *Docteur Jekyll & mister Hyde* et écrit par *Robert Louis Stevenson* (*cf.*, chapitre 2). Ici, la technique d'interprétation est similaire aux relations d'augmentations car elle consiste à analyser à la fois les descriptions éditoriales (*ex.*, champs de la zone 200) et les codes de fonction des responsabilités (*ex.*, champs 7XX$4). Toutefois, dans le cas des traductions, il faut également considérer les différentes langues qui peuvent être mentionnées. Dans l'exemple, la traduction est ainsi détectée grâce aux langues différentes dans la zone 101 (101$c = langue originale) ainsi qu'avec le code de fonction "traducteur" (730) d'une responsabilité secondaire présente dans le champ 702$4. L'autre dérivation, la relation l'adaptation, est ici détectée par le code de fonction "auteur original" (100) dans le champ 700$a ainsi que par le texte en langue naturelle du champ 200$e. Une particularité des dérivations se situe dans la manière de les modéliser, une fois interprétées. Par exemple, une traduction (*ex.*, anglais vers français) produit une nouvelle Expression d'une Œuvre existante alors qu'une adaptation (*ex.*, d'un livre en film) consiste en une relation entre deux Œuvres distinctes (l'adaptation étant alors une Œuvre avec son propre créateur).

Définition 4.2.4 (AGG : Relation d'agrégation)**.** . La métrique AGG détecte le pourcentage de notices décrivant une ou plusieurs relations d'agrégation. Une agrégation consiste en différents contenus intellectuels ou éditoriaux qui sont liés à une unique entité. Une agrégation peut exister au niveau intellectuel (*ex.*, collection d'Œuvres) ou au niveau éditorial (*ex.*, articles d'un journal).

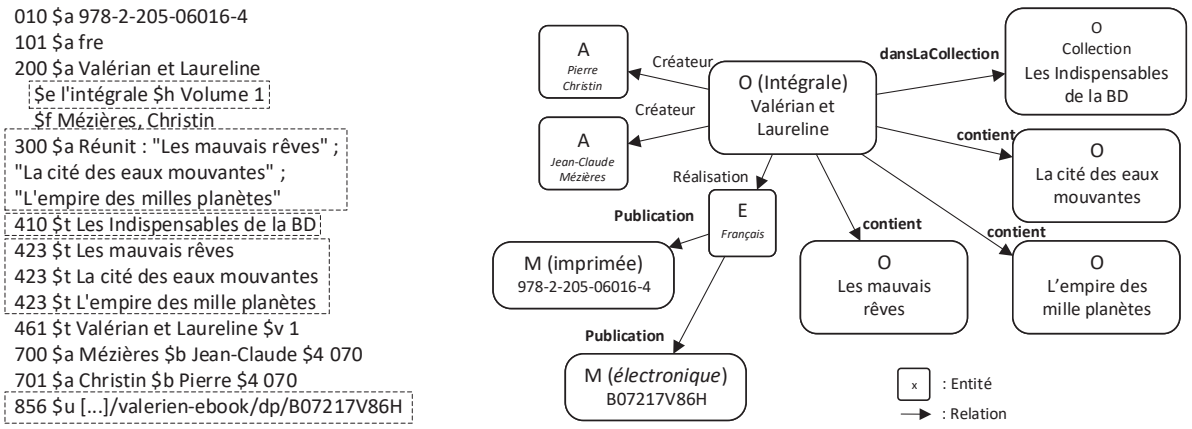
Pour comprendre cette notion, la figure 4.3 présente une notice décrivant une édition intégrale de la bande dessinée *Valérian et Lauréline* contenant plusieurs relations d'agrégations qui sont (1) une collection d'Œuvres, (2) une Œuvre d'agrégation et (3) des publications multiples. Dans le premier cas, une collection d'Œuvres est une agrégation éditoriale de plusieurs Œuvres. Dans la figure 4.3a, la collection est décrite dans le champ 410$*t* de la notice et peut donner lieu à une Œuvre à part entière appelée "*Les indispensables de la BD*" et liée à l'Œuvre "*Valérian et*



(a) Notice UNIMARC de dérivations (b) Exemple de modélisation FRBR

FIGURE 4.2 – Exemple d'interprétation de relations bibliographiques de dérivations

Lauréline". La seconde agrégation concerne trois Œuvres distinctes, décrites depuis les champs 423$\$t$ de la notice, qui sont groupées dans cette édition "intégrale". Dans notre exemple, figure 4.3b, cette agrégation peut être représentée comme une Œuvre qui est liée aux trois Œuvres par la relation *contient*. La dernière agrégation concerne des publications multiples d'une même Expression. Ici, l'édition (intégrale) en français a été publiée en deux formats, imprimée avec un ISBN (champ 010$\$a$) et numérique avec un lien (champ 856$\$u$). Cette dernière agrégation peut se traduire par deux manifestations distinctes d'une même Expression.



(a) Notice UNIMARC d'agrégations (b) Exemple de modélisation FRBR

FIGURE 4.3 – Exemple d'interprétation de relations bibliographiques d'agrégations

L'exemple d'interprétation des agrégations de la figure 4.3 illustre une particularité dans la détection de ces relations qui concerne le sens des liens qui doivent être interprétés, ces derniers pouvant être ascendants ou descendants. En effet, même si les entités représentant des créations originales, des collections ou des parties sont modélisées comme des Œuvres, il existe une hiérarchie implicite entre les entités "contenantes" ou "contenues" dans les normes documentaires. Il convient donc d'analyser quelles données font référence à quel niveau dans cette hiérarchie des agrégations. Dans notre exemple, figure 4.3b, le lien de la notice vers sa collection est un lien ascendant (la collection est un contenant) quand le lien de la notice vers les trois œuvres conte-

nues est un lien descendant (c'est l'Œuvre de la notice qui est un contenant). Nous notons que dans certains formats de notices comme MARC21, les liens (zone 4XX de UNIMARC) n'existent pas et l'interprétation des agrégations est bien souvent impossible ou très difficile car seulement renseignée dans des champs de notes (en texte plein) qui sont ajoutées par les documentalistes.

Définition 4.2.5 (COW : Relations de complémentarité)**.** . La métrique COW détecte le pourcentage de notices concernées par des relations complémentaires. Ces dernières sont établies entre deux Œuvres ayant une importance équivalente dans un contexte bibliographique défini.

La particularité des relations de complémentarité est que leur détection concerne nécessairement plusieurs Œuvres qui peuvent être dissimulées dans différentes notices bibliographiques sans lien explicite entre elles. L'intérêt de ces relations de complémentarité est d'offrir aux utilisateurs la capacité d'accéder facilement à l'ensemble d'une création originale si cette dernière a été décomposée en plusieurs Œuvres de manière intellectuelle ou éditoriale. La Figure 4.4 illustre cette notion avec deux notices, 4.4a et 4.4b, concernées par une relation de complémentarité.

010 $a 978-2-09-163866-9
101 $a fre
200 $a DCG epreuve 10
$e manuel et applications
$f Odile Barbe,
Laurent Didelot, Jean-Luc Siegwart
215 $a 1 volume (763 pages)
700 $a Barbe-Dandon $b Odile $4 070
700 $a Didelot $b Laurent $4 070
700 $a Siegwart $b Jean-Luc $4 070

(a) Notice UNIMARC 1

010 $a 978-2-09-163867-6
101 $a fre
200 $a DCG epreuve 10
$e corriges des applications
$f Odile Barbe,
Laurent Didelot, Laurence Cassio
215 $a 1 volume (271 pages)
700 $a Barbe-Dandon $b Odile $4 070
700 $a Didelot $b Laurent $4 070
700 $a Cassio $b Laurence $4 070

(b) Notice UNIMARC 2

FIGURE 4.4 – Exemple de notices bibliographiques complémentaires

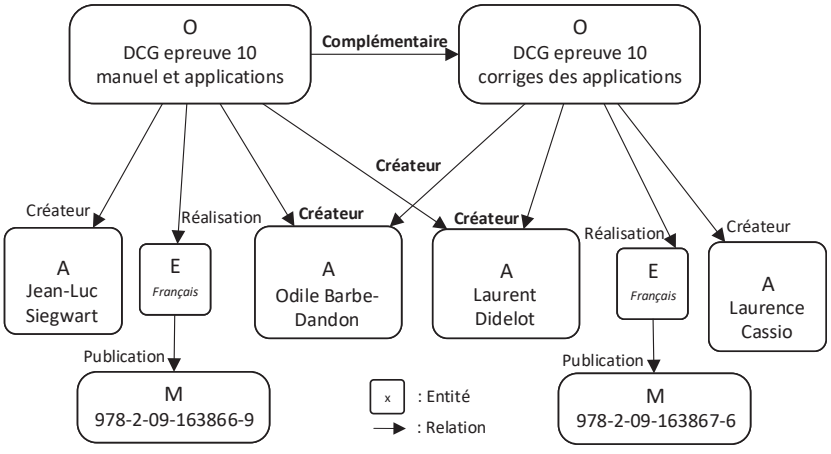


FIGURE 4.5 – Exemple de modélisation pour les notices complémentaires

Dans cet exemple, figure 4.4, il s'agit de ressources pédagogiques sur le domaine de la comptabilité où une des ressources contient des exercices préparatoires et l'autre ressource contient les corrigés correspondants. La modélisation de cette complémentarité, figure 4.5, se traduit par deux Œuvres liées ayant ici des responsabilités (créateurs) en commun. Ici, aucun lien n'est explicitement établi entre les champs des deux notices impliquant de devoir déduire la relation par un processus

spécifique (*ex.*, déduplication). Dans notre exemple, une analyse sur la similarité des titres et responsabilités des deux notices permet la détection de cette complémentarité. Toutefois, nous remarquons qu'une validation manuelle peut être nécessaire car la détection de cette relation n'implique pas qu'elle soit considérée comme "juste" par l'institution possédant les ressources.

4.2.2 Analyse des données requises pour l'interprétation

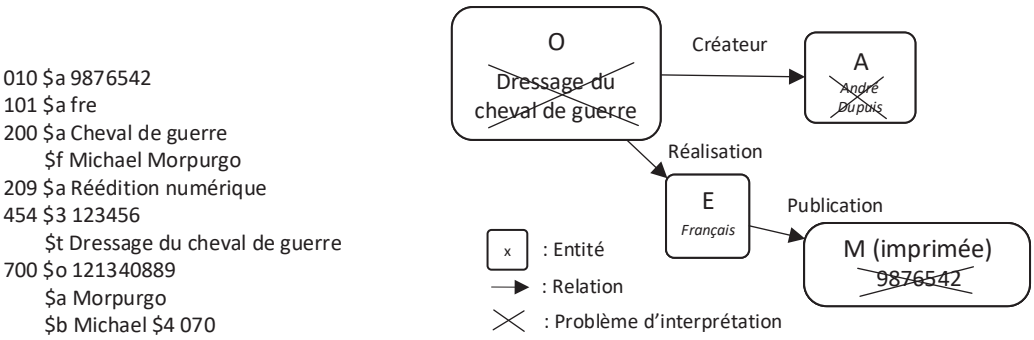
La qualité d'interprétation des connaissances bibliographiques varie en fonction des informations qui sont disponibles dans les données en entrée (*ex.*, dans les champs des notices). Ces informations peuvent effectivement être manquantes ou soumises à des pratiques de gestion spécifiques. Il peut être donc nécessaire, avant d'analyser les relations bibliographiques (décrites précédemment), d'évaluer quelles données sont disponibles afin d'adapter les critères d'interprétation ou corriger/compléter ces données. Dans ce contexte, le projet TelPlus [82] a défini six prérequis permettant d'évaluer la capacité d'interprétation de données bibliographiques initiales dans un processus de FRBRisation. Dans la suite, nous réutilisons ces pré-requis pour proposer dix métriques incluant quatre nouvelles métriques liées à des problèmes récurrents de catalogage. La table 4.1 liste et définit les six métriques du projet TelPLus avec un titre en français et une notation basée sur une notice r et un champ c. Ces métriques, respectivement nommées **MID**, **MPD**, **MUT**, **MOT**, **MRC** et **MAR** évaluent le pourcentage de notices ne respectant pas les prérequis définis. Certaines des erreurs/manquements considérés dans ces métriques concernent des problèmes cruciaux comme le manque de fonction de responsabilité d'un agent d'une ressource rendant plus difficile la distinction, par exemple, entre une dérivation et une augmentation.

Code	Problème / manquement	Exemple de notation
MID	Pas d'identifiant de notice	$\forall c \in r, c \not\rightsquigarrow$ 'identifiant de notice'
MPD	Pas de date de publication	$\forall c \in r, c \not\rightsquigarrow$ 'date de publication'
MUT	Pas de titre uniforme	$\forall c \in r, c \not\rightsquigarrow$ 'titre uniforme'
MOT	Pas de titre original	$\forall c \in r, c \not\rightsquigarrow$ 'titre original'
MRC	Pas de fonction de responsabilité	$\forall c \in r, c \not\rightsquigarrow$ 'code de fonction'
MAR	Pas de responsabilité	$\forall c \in r, c \not\rightsquigarrow$ 'responsabilité'

TABLE 4.1 – Métriques d'analyse des pré-requis à l'interprétation des notices dans TelPlus

Nos travaux sur l'interprétation des notices nous ont mené à considérer des cas supplémentaires et récurrents d'erreurs sur les données initiales ayant un impact sur la qualité d'interprétation des connaissances bibliographiques. Pour illustrer ces cas, la figure 4.6 présente un exemple de notice contenant des erreurs et pratiques qui sont différentes de celles énumérées dans TelPlus.

Dans cet exemple, figure 4.6a, on observe d'abord une absence d'information sur le format ou contenu de la publication. Cheval de Guerre, de Michael Morpurgo étant un roman plusieurs fois adapté (*ex.*, au cinéma), il est impossible de dire si la notice décrit une des adaptations ou la création originale. Ensuite, la zone 454 fait référence à une traduction de l'Oeuvre mais ici le lien vers l'autorité de titre originale est erronée et mène à la mauvaise Œuvre originale. De plus, l'identifiant de type ISNI de la responsabilité principale (zone 700) est incorrect et pointe vers une autre personne (André Dupuis). Enfin, une information sur la publication a été ajoutée de manière spécifique dans le champ 209, qui n'est pas conforme à la norme utilisée (ici UNIMARC). La conséquence de ces problèmes dans les données donne un résultat d'interprétation incorrect comme illustré sur la figure, 4.6b. A partir de ces observations, nous proposons quatre nouvelles métriques permettant d'anticiper ces différents problèmes que nous définissons ci-après.



(a) Notice UNIMARC incomplète (b) Modélisation erronée

FIGURE 4.6 – Problèmes d'interprétation à cause de données initiales manquantes ou erronées

Définition 4.2.6 (Métrique MTF : Missing Type and Form)**.** La métrique MTF évalue l'absence du type ou format matériel de publication.

Bien que cette information soit relative au domaine éditorial, elle peut servir à distinguer différentes Expressions d'une Œuvre ou éventuellement différentes Œuvres entre-elles. Les métriques **TLE** et **RLE** font références aux liens vers d'autres notices ou ressources qui sont erronées.

Définition 4.2.7 (Métrique TLE : Title Linkage Error)**.** La métrique TLE évalue la présence de liens erronés au niveau des autorités de titres.

Définition 4.2.8 (Métrique RLE : Responsibility Linkage Error)**.** La métrique RLE évalue les liens erronés des autorités de responsabilités (*ex.*, champ 700$3 pour un lien vers l'autorité auteur dans UNIMARC).

Définition 4.2.9 (Métrique CPN : Cataloguing Practices and Norms)**.** La métrique CPN évalue la présence d'autres pratiques spécifiques et normes ayant un impact sur l'interprétation des notices.

Le cas de cette quatrième métrique est plus spécifique aux pratiques de catalogage d'une institution. En effet, l'utilisation de standards comme l'International Standard Bibliographic Description (ISBD), qui préconise l'usage de ponctuations spécifiques pour délimiter différentes informations, ou encore l'usage de champs en dehors des normes (*ex.*, 209 dans l'exemple) augmente l'effort d'interprétation. Cette dernière métrique permet ici d'anticiper l'interprétation de ces spécificités et donc de réduire cet effort.

La table 4.2 résume la définition de ces quatre métriques avec une notation formelle où c_x désigne un champ dans une notice r et la fonction val_x retourne la valeur de ce champ.

4.2.3 Évaluation de la pertinence des règles pré-établies

La dernière partie de cette section concerne le cas où un jeu de règles d'interprétation de données bibliographiques est déjà défini. La phase d'évaluation de ce jeu de règles, qui est assez peu évoquée dans la littérature du domaine, peut avoir un impact significatif sur les efforts d'interprétation et le coût global d'un projet d'intégration de données. Dans ce contexte, nous proposons trois métriques dédiées à l'évaluation d'un jeu de règles d'interprétation prédéfini.

Définition 4.2.10 (Métrique MR : Missing Rules)**.** La métrique MR évalue les règles qui sont manquantes, dans un jeu de règles donné, pour l'interprétation de données bibliographiques.

Code	Problème / manquement	Notation
MTF	Pas de format d'édition	$\forall c \in r, c \not\rightsquigarrow$ 'format'
TLE	Erreur de lien titre	$c \in r, c \rightsquigarrow$ 'titre' $\wedge \nexists r_2 \in \mathcal{R}$ $(c_2 \in r_2 \wedge c_2 \rightsquigarrow$ 'identifiant de notice' $\wedge\ val_{c_2} = val_c)$
RLE	Erreur de lien responsabilité	$c \in r, c \rightsquigarrow$ 'responsabilité' $\wedge \nexists r_2 \in \mathcal{R}$ $(c_2 \in r_2 \wedge c_2 \rightsquigarrow$ 'identifiant de notice' $\wedge\ val_{c_2} = val_c)$
CPN	Pratiques de catalogage	$r \hookrightarrow$ 'règle de catalogage' $\vee\ r \hookrightarrow$ 'ponctuation' $\vee$ $(c \in r, c \rightsquigarrow$ 'données locales' $\vee\ \varphi(val_c) \neq \emptyset)$

TABLE 4.2 – Nouvelles métriques d'analyse des pré-requis pour l'interprétation des notices

La métrique MR peut servir à détecter les champs de notices bibliographiques qui ne sont considérés par aucune règle prédéfinie, causant ainsi une perte d'information dans un processus de migration. Par exemple, elle peut être utilisée pour vérifier si un jeu de règles considère les codes de fonctions des auteurs dans les notices bibliographiques afin de distinguer les narrateurs, des traducteurs ou autres fonctions. Il est également possible de combiner cette métrique MR avec des règles d'interprétation des relations bibliographiques avancées (*ex.*, MR-AUG pour mesurer l'absence de règles d'interprétation des relations augmentations).

Définition 4.2.11 (Métrique UR : Unused Rules)**.** La métrique UR mesure les règles prédéfinies qui ne sont pas utiles pour un ensemble de données bibliographiques à intégrer.

La métrique UR peut s'avérer utile pour améliorer les performances d'un outil ne proposant pas d'optimisation sur l'application des règles aux champs des notices. Par exemple, un ensemble de règles de migration qui contiennent des correspondances avec des champs qui ne sont pas utilisés génère des calculs inutiles qui peuvent être coûteux sur de grands volumes de données.

Définition 4.2.12 (Métrique CR : Conflicting Rules)**.** La métrique CR évalue les règles conflictuelles.

Des conflits en règles peuvent subvenir quand l'interprétation de certains champs mènent à des contradictions dans les données produites. Par exemple, il peut avoir été décidé d'attacher le résumé d'un livre à la classe Œuvre dans un projet d'intégration de données et sur la classe Expression dans un autre projet. Il peut être fréquent d'avoir des règles en conflits si différents experts travaillent sur un même projet de migration. En conséquence, il est crucial de pouvoir anticiper l'existence de ces conflits.

En résumé, les métriques CORE, AUG, DER, AGG, COW, MID, MPD, MUT, MOT, MRC, MAR, MTF, TLE, RLE, CPN, MR, UR et CR présentées précédemment, sont dédiées à l'évaluation de l'interprétation de données bibliographiques en amont d'un processus d'intégration. Dans la section suivante, nous décrivons des métriques dédiées à l'évaluation de ce processus.

4.3 Évaluer la transformation de métadonnées bibliographiques

La phase de transformation, c'est à dire la phase automatisée d'un processus d'intégration de données, consiste en l'application de règles pré-établies sur les données en entrée du processus afin de réaliser leur intégration dans une base de connaissances cible. Les métriques que nous présentons dans cette section sont relatives aux performances du processus car cela peut avoir un

impact sur la faisabilité d'un projet d'intégration, par exemple lorsque différentes itérations sont nécessaires. Ces métriques ont donc pour objectif de faciliter l'évaluation des performances de l'outil utilisé pour l'intégration automatique de données bibliographiques. Bien que ces dernières peuvent être difficiles à implémenter pour certains outils existants, elles peuvent néanmoins influencer le développement des futures solutions d'intégration (*ex.*, de FRBRisation).

Définition 4.3.1 (Métrique ETC : Execution Time Cost of the whole extraction)**.** La métrique ETC mesure le temps nécessaire à un outil pour appliquer l'ensemble des règles de migration à l'ensemble des notices d'un catalogue donné.

Dans le cas d'un processus d'intégration de données dit "à la volée", c'est à dire en temps réel sur un système d'information bibliographique, cette métrique permet la comparaison d'outils, à qualité similaire, en termes de performances du système pendant l'intégration.

Définition 4.3.2 (Métrique ETD : Execution Time for Deduplication)**.** La métrique **ETD** mesure le temps d'exécution nécessaire au processus de déduplication des entités à intégrer.

Cette métrique distingue le temps d'exécution de la déduplication du reste du processus car nous avons vu au chapitre 3 que cette étape pouvait être coûteuse selon la taille et la complexité des données en entrée [28]. C'est le cas notamment si la déduplication est effectuée après l'application des règles sur les données initiales, résultant en un grand nombre d'entités à comparer [82].

La Table 4.3 synthétise ces deux métriques. Nous admettons que leur implémentation est fortement dépendante de l'outil concerné. Toutefois, elles peuvent devenir des critères essentielles pour la réalisation de futures outils d'intégration.

Métrique	Définition
ETC	Temps d'execution pour intégrer de nouvelles entités et relations
ETD	Temps de déduplication

TABLE 4.3 – Métriques pour l'évaluation du processus de migration des données

4.4 Évaluer la qualité des métadonnées bibliographiques

Dans la littérature du domaine documentaire, l'évaluation *a posteriori* d'un processus d'intégration de données bibliographiques se limite généralement au calcul de mesures de complétudes sur les données intégrées ou de satisfaction des utilisateurs [20, 127]. Notre observation est que les métriques et résultats proposés ne tiennent pas toujours compte des spécificités du domaine bibliographique [5, 105]. C'est pourquoi, dans cette dernière partie, nous présentons neuf métriques permettant d'évaluer la qualité d'une base de connaissances bibliographiques qui a été créée à l'issue d'un processus de migration de notices. Nous répartissons les métriques en deux catégories qui sont respectivement liées à l'évaluation des instances migrées et l'évaluation de la sémantique de ces instances. Si la première catégorie contient des métriques plutôt communes aux projets existants d'intégration de données, la seconde contient des métriques intégrant la notion spécifique de motifs de connaissances bibliographiques.

4.4.1 Méthode d'évaluation

Nous considérons qu'une évaluation sur l'intégration de motifs de connaissances est bien souvent subjective car un processus d'intégration automatique répond à une méthodologie d'interprétation et de modélisation des connaissances qui peut-être spécifique à un projet. C'est pourquoi ce

type d'évaluation doit être réalisé selon un référentiel expert qui est défini préalablement dans le cadre du projet. Dans ce contexte, les métriques que nous proposons dans cette section comparent systématiquement une base de connaissances à évaluer $\mathcal{T}$ avec une base de connaissances experte $\mathcal{E}$ définie préalablement. Les métriques présentées ci-après peuvent être utilisées pour améliorer les outils ou leurs règles qui sont dédiés à la création de la future base de connaissances. Dans cet objectif, les métriques sont orientées pour calculer des taux d'erreurs, c'est à dire que plus le résultat est proche de 100% plus il y a d'erreurs dans la base de connaissances et plus il est nécessaire de réaliser des ajustements au niveau de l'outil en question et/ou de ses règles.

Pour chacune des deux catégories de métriques, nous proposons un tableau de synthèse incluant un exemple de notation en logique du premier ordre. Chaque notation est basée sur le principe de comparaison de deux bases de connaissances, $\mathcal{T}$ qui correspond à la base à évaluer et $\mathcal{E}$ qui correspond à la base experte. Les comparaisons sont réalisées en tenant compte du type de données à comparer. Étant données $e \in \mathcal{E}$ et $t \in \mathcal{T}$, les entités des deux bases sont comparées en fonction de leur *type* et de leur *valeur* textuelle principale (*i.e.*, label préféré) :

$$e \equiv t \iff type_e = type_t \wedge valeur_e = valeur_t$$

Pour les relations, c'est à dire les propriétés qui lient deux entités, la comparaison est effectuée selon le *type* de relation et selon les *entités* qui sont liées. Soit $e \in \mathcal{E}$ et $t \in \mathcal{T}$, désignant des relations et $entité^1$ et $entité^2$ les entités aux extrémités de ces relations, alors :

$$e \equiv t \iff type_e = type_t \wedge entité^1_e = entité^1_t \wedge entité^2_e = entité^2_t$$

Les attributs des entités sont comparés selon leur *type*, leur d'*entité* d'origine et leur *valeur* :

$$e \equiv t \iff type_e = type_t \wedge entité_e = entité_t \wedge valeur_e = valeur_t$$

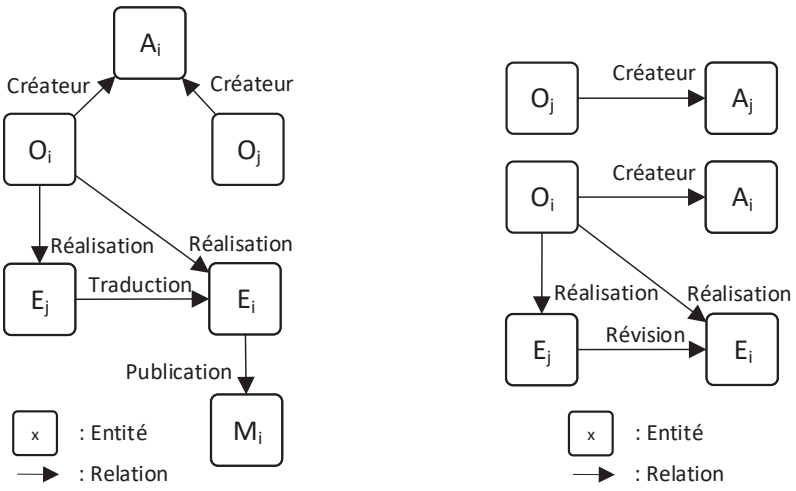
4.4.2 Mesure sur les instances

Dans cette partie, nous présentons des métriques dédiées à l'évaluation des instances d'une nouvelle base de connaissances (incluant les propriétés et les relations des entités). La Figure 4.7 illustre ce type d'évaluation en montrant un exemple de deux bases de connaissances à comparer.

Observations sur l'exemple. La base experte $\mathcal{E}$, figure 4.7a, présente deux Œuvres bibliographiques O_i et O_j, créées par le même auteur A_i. La première Œuvre O_i a été réalisée en deux Expressions E_i et E_j où E_i est issue d'une traduction de E_j et a donné lieu à une publication M_i. Dans la base à évaluer $\mathcal{T}$, figure 4.7b, nous observons qu'il manque les informations relatives à la publication M_i. De plus, les deux Œuvres ne sont pas liées au même auteur A_i, conséquence d'une erreur de déduplication. Enfin, la propriété symbolisant la traduction a, dans cet exemple, une sémantique incorrecte (*ex.*, liée à une erreur d'interprétation) et a pour valeur "*Révision*".

Dans le contexte de l'évaluation des différences entre les instances de la base à évaluer et de la base experte, illustrées par les précédentes observations, nous proposons quatre métriques dont les codes respectifs sont **MD**, **IAD**, **SMD** et **DLE**.

Définition 4.4.1 (Métrique MD : Missing Data)**.** La métrique MD calcule les données manquantes, c'est à dire les données qui sont présentes dans la base experte $\mathcal{E}$ mais absentes de la base à évaluer $\mathcal{T}$.



(a) Base de connaissance experte $\mathcal{E}$ (b) Base de connaissance à évaluer $\mathcal{T}$

FIGURE 4.7 – Exemples de différences structurelles au niveau des instances d'une base de connaissances à évaluer $\mathcal{T}$ et celles d'une base experte $\mathcal{E}$

Le calcul de la métrique MD retourne le pourcentage de données absentes dans $\mathcal{T}$ par rapport aux données totales dans $\mathcal{E}$. Cette métrique peut être également spécialisée pour chaque type de données (*ex.*, MD-E pour les entités, MD-R pour les relations).

Définition 4.4.2 (Métrique IAD : Incorrectly Added Data)**.** La métrique IAD correspond aux données qui ont été incorrectement produites par le processus d'intégration. Ces données peuvent être des doublons ou des données erronées dans la base de connaissances cible.

Cette métrique permet de détecter deux problèmes dans le processus d'intégration qui sont des erreurs de déduplication ou des incohérences dans les règles de transformation. IAD est calculée comme le nombre de données dans $\mathcal{T}$ qui ne sont pas présentes dans $\mathcal{E}$, le tout divisé par le nombre total de données produites dans $\mathcal{T}$. Comme pour la mesure précédente, IAD peut être déclinée selon le type de données (*ex.*, IAD-E pour les entités, IAD-R pour les relations).

Définition 4.4.3 (Métrique SMD : Semantic Mismatch Data)**.** La métrique SMD évalue les incohérences sémantiques entre les propriétés et relations des bases $\mathcal{T}$ et $\mathcal{E}$. C'est à dire, pour une information intégrée depuis des données initiales, la relation ou propriété produite dans la base $\mathcal{T}$ a une sémantique différente de celle issue des mêmes données dans le catalogue expert $\mathcal{E}$.

La notion d'incohérence sémantique (*cf.*, *Semantic Gap*[1]) désigne ici deux parties d'un modèle (de bases de connaissances) dont la structure est identique mais dont le sens attribué aux propriétés et/ou relations est différent. La métrique compare donc le résultat du taux d'incohérences sémantiques, dans $\mathcal{T}$ par rapport à $\mathcal{E}$, au nombre de données dans $\mathcal{T}$. Par exemple, SMD permet de détecter qu'une relation *traduit par*, dans la base experte $\mathcal{E}$, se retrouve être *contributeur* dans la base $\mathcal{T}$, ce qui n'est ici pas suffisant pour la compréhension de cette connaissance dans $\mathcal{T}$.

Définition 4.4.4 (Métrique DLE : Data Linkage Error)**.** La métrique DLE permet de détecter les liens erronés entre des entités ou propriétés de la base à évaluer $\mathcal{T}$ et des données issues de sources externes de type *référentiel spécifique* ou *web de données*. Cette métrique peut être appliquée sans utiliser la base experte $\mathcal{E}$.

1. `https://en.wikipedia.org/wiki/Semantic_gap`

Cette métrique calcule le nombre de liens erronés dans $\mathcal{T}$, divisés par le nombre total de liens représentés dans $\mathcal{T}$. L'intégration de données impliquant souvent des enrichissements de multiples sources, il peut être nécessaire d'évaluer si les liens ajoutés ou construits sont correctes ou non.

La table 4.4 présente une synthèse des quatre métriques décrites jusqu'à présent et propose, pour chacune d'elle, un exemple de notation formelle. La notation relative à la métrique ME montre que la donnée e, de la base experte $\mathcal{E}$, n'a pas d'équivalent dans la base à évaluer $\mathcal{T}$. Le calcul de cette métrique retourne donc le rapport entre le nombre de données manquantes et le nombre de données totales dans la base experte. Nous notons que cette métrique peut être spécialisées à chaque type de données à évaluer définissant ainsi les métriques *MD-E* pour les entités, *MD-R*, pour les relations et *MD-P* pour les propriétés. La métrique IAD est calculée comme le nombre incorrecte de données dans $\mathcal{T}$ que l'on divise par le nombre total de données dans $\mathcal{T}$. Comme pour MD, la métrique IAD peut être spécialisée selon les différents types de données à évaluer. Le calcul de la métrique SMD dépend du nombre de données, ayant une sémantique différente entre les deux bases, que l'on compare au nombre de données totale. Un exemple de différence sémantique est l'usage d'une propriété, dans la collection évaluée, dont la sémantique est plus large que la propriété attendue dans la collection experte (ex., *véhicule* et *voiture*) ou est une subsomption [2] de cette propriété notée $t \subset e$). Enfin, la métrique DLE est calculée en comparant le nombre de liens erronés dans $\mathcal{T}$ avec le nombre total de liens dans $\mathcal{T}$.

Metric	**Related issue**	**Formal notation**
MD	Données manquantes	$e \in \mathcal{E}, \forall t \in \mathcal{T}, t \not\equiv e$
IAD	Données incorrectes	$t \in \mathcal{T}, \forall e \in \mathcal{E}, e \not\equiv t$
SMD	Problème de sémantique	$e \in \mathcal{E}, t \in \mathcal{T}, (t \subset e) \vee (e \subset t)$
DLE	Liens externes erronés	$t \in \mathcal{T}, t \rightsquigarrow$ 'external link' $\wedge$ ($\exists e \in \mathcal{E} \wedge e \rightsquigarrow$ 'external link' $\wedge valeur_t \neq valeur_e \wedge source_t = source_e$)

TABLE 4.4 – Synthèse des métriques d'évaluation sur les instances intégrées

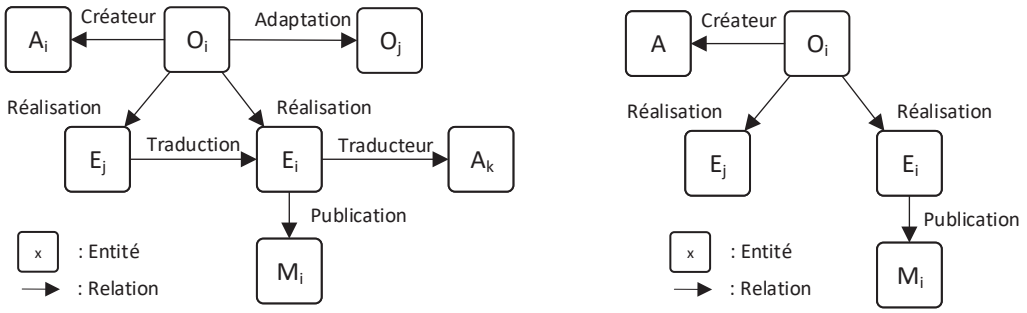
4.4.3 Mesures sur les motifs de connaissances

Les métriques présentées dans cette catégorie évaluent l'intégration de motifs de connaissances bibliographiques dans un processus d'intégration de données. La figure 4.8 illustre l'évaluation de cette intégration avec un exemple contenant une base experte et une base à évaluer.

Observations sur l'exemple. Ici, la base experte $\mathcal{E}$ décrit deux Œuvres O_i et O_j où O_j est une adaptation de O_i et O_i a été réalisée en deux Expressions E_i et E_j. E_i est une traduction de E_j par l'Agent A_k et a donné lieu à une publication M_i. Dans la base $\mathcal{T}$ à évaluer, nous constatons que l'entité (O_j) et la relation décrivant l'adaptation de O_i n'a pas été intégrée. Dans le motif de traduction, les deux Expressions E_i et E_j ont bien été intégrées mais la relation de traduction entre ces deux entités n'a pas été créée. De plus le traducteur A_k n'a pas été intégré. Cet exemple montre que certains motifs de connaissances peuvent être intégrés partiellement.

Pour évaluer l'intégration des motifs de connaissances bibliographiques, nous considérons qu'un motif peut avoir été partiellement intégré et qu'il est utile de comprendre qu'elle aspect du

2. https://en.wikipedia.org/wiki/Hierarchy#Subsumptive_containment_hierarchy



(a) Base de connaissances experte $\mathcal{E}$ (b) Base de connaissances à évaluer $\mathcal{T}$

FIGURE 4.8 – Différences entre les motifs de connaissances d'une base à évaluer $\mathcal{T}$ et ceux d'une base experte $\mathcal{E}$

motif est manquant afin de mieux comprendre les faiblesses de l'outil évalué. Nous considérons trois niveaux dans un motif de connaissances, l'*entité principale*, la *relation principale* et les *éléments secondaires*. Par exemple, dans un motif de traduction d'une Œuvre, l'entité principale est l'Expression traduite de cette Œuvre, la relation principale est la relation entre l'Expression originale de l'Oeuvre et l'Expression traduite et enfin, l'Agent traducteur est un élément secondaire.

Nous définissons quatre nouvelles métriques relatives aux différents niveaux d'intégration d'un motif. Nous proposons ci-après des notations formelles pour la détection de ces niveaux. Pour cela, nous introduisons l'ensemble $\mathcal{E}' \in \mathcal{E}$ qui inclut tous les éléments principaux d'un motif, c'est à dire l'entité principale et la relation principale, ainsi que l'ensemble $\mathcal{E}'' \in \mathcal{E}$ qui contient tous les éléments secondaires du motif. Nous définissons maintenant les quatre métriques de cette partie. Pour rappel, les métriques suivantes utilisent le principe de comparaison des données entre une base de connaissances experte $\mathcal{E}$ et une base à évaluer $\mathcal{T}$.

Définition 4.4.5 (Métrique MEND : Main Entity Not Detected)**.** La métrique MEND évalue les problèmes de détection des entités principales d'un motif.

Pour chaque motif à considérer, l'entité principale d'un motif doit être définie par les experts et considérée par la métrique. Par exemple, dans un motif de traduction, l'Expression contenant la langue traduite peut être considérée comme l'entité principale.

Définition 4.4.6 (Métrique MRND : Main Relationship Not Detected)**.** La métrique MRND évalue si la relation principale d'un motif, liée nécessairement à l'entité principale, est correctement identifiée ou non.

MRND permet de calculer le pourcentage de relations principales (d'un motif) qui n'ont pas été détectées par rapport à l'ensemble des relations principales de la base à évaluer. Ces relations sont également définies préalablement par les experts pour chaque motif à intégrer. Dans l'exemple d'un motif de traduction, la métrique MRND permet de détecter (dans le cas où l'entité principale, *Expression traduite*, est identifiée) s'il existe bien une relation entre l'Expression originale et l'entité principale avec une sémantique correcte (*ex.*, *traduction de*).

Définition 4.4.7 (Métrique ESE : Error(s) in Secondary Elements)**.** La métrique ESE évalue les erreurs dans l'intégration des éléments secondaires d'un motif.

Les éléments secondaires sont toutes les entités et relations d'un motif, autres que l'entité et la relation principale, qui sont définies par les experts comme ayant une importance dans la modélisation d'un motif en question. Pour le calcul de ESE, nous considérons que l'entité principale et

la relation principale du motif ont été correctement intégrées mais que ces éléments secondaires (*ex.*, *le traducteur* dans un motif de traduction) sont manquants ou incorrectement intégrés. La métrique ESE calcule alors un pourcentage d'éléments secondaires qui ne sont pas intégrés par rapport à l'ensemble des éléments secondaires intégrés.

Nous proposons également la métrique *FPND* qui agrège ces trois métriques.

Définition 4.4.8 (Métrique FPND : Full Pattern Not Detected)**.** FPND évalue l'équivalence des motifs de connaissances complets entre la base experte $\mathcal{E}$ et la base à évaluer $\mathcal{T}$. La complétude d'un motif dépend de la présence de l'entité principale, la relation principale ainsi que tous les éléments secondaires.

Cette métrique calcule le pourcentage de motifs de connaissances dans $\mathcal{E}$ qui ne sont pas reproduits à l'identique dans $\mathcal{T}$. Par exemple, un motif de traduction d'une Œuvre est détecté entièrement quand une nouvelle Expression (traduction) est associée à l'Œuvre originale et qu'il existe une relation entre l'Expression originale et l'Expression traduite ainsi qu'une relation secondaire entre l'Expression traduite et les différents Agents traducteurs détectés. Cette métrique est par définition très stricte et a pour unique objectif de livrer une vision globale des possibilités d'un outil ou de règles d'intégration de connaissances bibliographiques.

La table 4.5 présente une synthèse des quatre métriques décrites pour cette seconde catégorie et propose, pour chacune d'elle, un exemple de notation formelle. Dans la notation formelle proposée, nous considérons qu'une entité principale e' doit avoir une entité équivalente t dans la base de connaissances à évaluer. De plus, nous remarquons qu'avec cette notation générique, les métriques MRND and ESE ont une expression similaire.

Métrique	**Problème**	**Notation formelle**
FPND	Motif entier non intégré	$e' \in \mathcal{E}', e'' \in \mathcal{E}'', \forall t \in \mathcal{T}, \forall t' \in \mathcal{T}, t \not\equiv e' \vee t' \not\equiv e''$
MEND	Entité principale non intégrée	$e' \in \mathcal{E}', \forall t \in \mathcal{T}, t \not\equiv e'$
MRND	Relation principale non intégrée	$e' \in \mathcal{E}', \forall t \in \mathcal{T}, t \not\equiv e'$
ESE	Erreurs des éléments secondaires	$e'' \in \mathcal{E}'', \forall t \in \mathcal{T}, t \not\equiv e''$

TABLE 4.5 – Synthèse des métriques d'évaluation sur les motifs de connaissances intégrées

Nous proposons une métrique supplémentaire qui combine les cinq métriques principales (dans les deux catégories précédentes) liées à l'évaluation de la qualité du processus de FRBRisation (à savoir MD, IAD, DLE, SMD et FPND). Cette métrique globale, appelée OQF (*Overall Quality of FRBRization*), est définie par :

$$\frac{\alpha_1 MD + \alpha_2 IAD + \alpha_3 DLE + \alpha_4 SMD + \alpha_5 FPND}{\alpha_1 + \alpha_2 + \alpha_3 + \alpha_4 + \alpha_5}$$

où α_1, α_2, α_3, α_4 et α_5 sont des poids dont la somme est égale à 1.

Chaque poids indique l'importance accordée au problème de qualité évalué (parmi les cinq utilisés) pour le calcul de la qualité globale de la transformation. Un poid peut être fixé à 0 si le problème ne doit pas être considéré dans le calcul. Cette métrique OQF calcul un score de qualité entre 0 et 1 où un score minimal indique que le processus de FRBRisation n'a abouti à aucune erreur de qualité. Ce score reste néanmoins global et ne permet pas une évaluation fine de la qualité d'un outil de FRBRisation.

4.5 Benchmark BIB-R

Dans cette section, nous présentons notre benchmark BIB-R [3] pour l'évaluation des solutions de migration de métadonnées bibliographiques. Ce benchmark intègre l'ensemble des métriques présentées dans ce chapitre et propose également des jeux de données permettant d'évaluer des outils de FRBRisation. La particularité de BIB-R est qu'il considère les spécificités du domaine bibliographique en intégrant notamment des tests sur les erreurs typiques qui concernent les notices bibliographiques. Nous présentons différentes expérimentations, réalisées avec BIB-R, pour mieux comprendre les enjeux techniques liés au développement des outils de FRBRisation.

4.5.1 Jeux de données

Nous présentons ici deux jeux de données que nous avons crées manuellement puis validés par des experts et enfin mis à disposition de la communauté [4] pour l'évaluation des outils de migration de notices documentaires. Dans notre contexte, ces jeux de données sont des collections de notices. Le premier jeu de données est nommé **T42** (pour *42 tests*) et couvre les caractéristiques des métadonnées bibliographiques détaillées dans le chapitre 4. Le second jeu de données est nommé **BIB-RCAT** ("*Basically a Real-world CATalog*") et simule un sous-ensemble caractéristique d'un catalogue bibliographique du monde réel. Le jeu de données T42 est utilisé pour évaluer dans le détail les spécificités qui sont prises en compte ou non par un outil de migration quand le jeu de données BIB-RCAT livre une vision globale des possibilités d'un outil en conditions réelles.

Pour les deux jeux de données, nous mettons à disposition les notices dans deux formats MARC courants qui sont MARC21 et UNIMARC et nous fournissons une collection experte, réalisée manuellement, d'entités et de propriétés FRBR, correspondant à une migration idéale des notices en question. Cette collection experte permet notamment d'intégrer des tests liés aux métriques de qualité de la section 4 du chapitre 4. L'ensemble des données relatives à ces deux jeux de données sont accessibles librement à l'adresse : `http://bib-r.github.io/`.

Jeu de données T42

L'objectif de T42 est de permettre de savoir si un outil de migration tient compte des différentes spécificités du domaine bibliographique. Dans les sections précédentes, nous avons montré que ces spécificités s'articulent autour de cinq motifs de connaissances (CORE, AUG, DER, AGG et COW) et de particularités liées à un catalogue donné (ex., champs manquants, pratiques de catalogage). C'est pourquoi, les 42 tests de T42 s'articulent autour de ces cinq motifs et d'une sélection de métriques importantes à l'interprétation des notices. Nous définissons la notion de *test*, dans T42, comme la tâche de détection d'un motif auquel nous associons au plus une anomalie. Dans le contexte de nos travaux, nous choisissons de limiter chaque test à une seule anomalie au maximum pour éviter de complexifier le calcul et les statistiques associées au test. Il pourrait cependant être intéressant d'associer des combinaisons d'anomalies particulières à un test afin d'obtenir un niveau de lecture plus fin sur les problèmes d'un outil. Nous ajoutons qu'une anomalie qui est associée à un test n'empêche pas la détection du motif à condition que l'outil évalué intègre, pour certains tests, des fonctionnalités de nettoyage et de préparation des données en amont de la migration. Le tableau 4.6 présente une synthèse des différents tests de T42 avec leurs caractéristiques spécifiques.

Dans le tableau, le motif *Core* (à l'horizontal) désigne le motif élémentaire décrit nativement par chaque notice d'une collection et qui n'est pas un motif avancé. L'anomalie (à la verticale) nom-

3. `http://bib-r.github.io/`
4. `https://github.com/bib-r/bib-r/zipball/master`

	Core (1.x)	**AUG (2.x)**	**DER (3.x)**	**AGG (4.x)**	**COW (5.x)**
Basique	1.0	2.0	3.0	4.0	5.0
Pas de date de publication (MPD)	1.1	2.1	3.1	4.1	5.1
Pas d'identifiant de notice (MID)	1.2	2.2	3.2	4.2	5.2
Pas de format de ressource (MTF)	1.3	2.3	3.3	4.3	5.3
Lien erroné des autorités de titres (TLE)	1.4	-	-	-	-
Pas de titre uniforme (MUT)	1.5	2.5	-	4.5	5.5
Pas de titre original (MOT)	-	-	3.6	-	-
Lien erroné des responsabilités (RLE)	1.7	-	-	-	-
Pas de fonction de responsabilité (MRC)	1.8	2.8	3.8	4.8	5.8
Pas de responsabilité (MAR)	1.9	2.9	3.9	4.9	5.9
Pratiques de catalogage (CPN)	1.10	2.10	3.10	4.10	5.10

TABLE 4.6 – Principales caractéristiques des tests de T42

mée *Basique* correspond à une notice bien formée et ne comportant aucun problème nécessitant une technique d'interprétation spécifique. Ainsi, le *test 1.0* consiste à migrer des notices sans motif particulier et sans aucune erreur. Comme autre exemple, le *test 3.9* consiste à interpréter un motif de dérivation dans des notices ne contenant pas de description des responsabilités. Nous faisons remarquer que certains tests ne sont appliqués sur certains motifs car ils sont spécifiques à ce motif (*ex.*, 3.6, liés aux traductions dans le motif des dérivations) ou au contraire qu'ils ne sont pas pertinents vis à vis des notices proposées dans les catégories de motifs.

Jeu de données BIB-RCAT

Le jeu de données BIB-RCAT simule un catalogue de notices réelles contenant une multitude de motifs et d'anomalies à interpréter. Les notices sont issues de plusieurs catalogues du monde réel et ont été sélectionnées pour représenter l'ensemble des caractéristiques des métadonnées bibliographiques. C'est pourquoi, à la différence de T42, BIB-RCAT peut contenir des notices décrivant plusieurs motifs et contenant plusieurs anomalies. Ce jeu de données contient 560 notices et n'est conçu que pour donner un aperçu des possibilités d'un outil évalué et non ses performances en termes de temps de traitement.

Le tableau 4.7 présente les statistiques relatives aux deux jeux de données présentés, T42 (à gauche) et BIB-RCAT (à droite). Par exemple, le tableau indique que T42 inclut des notices décrivant des ressources documentaires dans trois langues différentes (anglais, français et allemand) et selon huit formats de document (*ex.*, livre, film, audio). Dans BIB-RCAT, la collection experte contiens 1922 entités à détecter avec au moins 9500 propriétés.

4.5.2 Évaluations de 3 solutions de migration

Nous présentons ci-après des expérimentations réalisées sur trois outils de l'état de l'art que nous avons sélectionné. Ces outils, qui sont librement accessibles, reposent sur des règles pour réaliser le processus de FRBRisation : (1) Variations FRBR [5] (**VFRBR**), (2) Extensible Catalog [6] (**XC**), (3) **FRBR-ML** [7]. Les calculs des résultats des métriques du Benchmark sur ces outils ont été réa-

5. VFRBR : `https://github.com/naimdjon/vfrbr-frbrize-marc`
6. XC : `http://www.extensiblecatalog.org/`
7. FRBR-ML : `https://github.com/naimdjon/marc2frbr`

Feature	T42	BIB-RCAT
Nombre de tests	42	-
Nombre de collections de notices	126	3
Nombre de langues de ressources	3	1
Nombre de type de ressources	8	4
Moyenne des notices (MARC)	10/test	560
Moyenne des champs / notice	18	17
Moyenne des entities (FRBR) à détecter	73/test	1922
Moyenne des propriétés (FRBR) à détecter	241/test	9517

TABLE 4.7 – Statistiques des jeux de données T42 et BIB-RCAT

lisés par un système automatique puis par une vérification manuelle systématique par plusieurs experts. Cette double validation nous a notamment permis de prendre certains décisions sur des ambiguïtés levées par le système automatique car les outils intègrent des interprétations différentes des modèles de notices MARC et du modèle conceptuel FRBR (ce qui impacte les résultats de métriques comme IAD ou SMD qui sont liées à l'intégration des données et leurs sémantiques).

Les expérimentations présentées ont pour objectif de vérifier l'hypothèse principale de cette thèse. C'est à dire que nous souhaitons observer l'impact de la prise en compte des spécificités bibliographiques sur la qualité produite par des outils de migration et d'enrichissement de métadonnées documentaires. Nous faisons remarquer que, pour les trois outils évalués, nous utilisons les modèles de règles initiaux qui sont fournis avec les outils. Ces outils sont donc évalués en l'état, c'est à dire tels quels depuis leur téléchargement. Leurs modèles de règles ne sont donc pas adaptés pour se conformer aux jeux de données du benchmark. En conséquence, nous nous attendons à de forts taux d'erreurs dans la détection de certaines relations bibliographiques avancées. Cependant, notre objectif n'est pas de réaliser une critique des outils, dont les règles initiales ne sont pas forcément représentatives du potentiel de ces solutions, mais bien d'étudier les possibilités offertes par l'intégration, ensuite, de nos métriques d'interprétation des connaissances que nous avons présenté dans ce chapitre, pour améliorer la qualité globale du processus de migration.

Prise en compte des motifs de connaissances

Dans une première expérimentation expérimentation, les trois solutions sont évaluées sur le jeu de données T42 selon 3 métriques, MEND (taux d'entités principales qui ne sont pas détectées dans un motif), MRND (taux de relations principales qui ne sont pas détectées dans un motif) et MD (taux de données globales qui ne sont pas migrées). La figure 4.9 présente les résultats de cette expérimentation avec l'outil VFRBR. Dans ce diagramme, les valeurs exprimées sont des erreurs. Plus le résultat est proche de 100%, plus il y a d'erreurs d'interprétation des motifs.

Le diagramme montre les résultats des trois métriques sur l'ensemble des tests de T42. Pour des raisons de lisibilité, MEND est représentée par des barres quand les deux autres métriques MRND et MD sont représentées par des courbes. Dans cette expérimentation de la solution VFRBR, les taux d'erreurs dans la détection des motifs sont globalement proches de 100% et les taux de données non migrées ne descendent pas en dessous de 40% d'erreur. Le problème dans ce contexte réside dans certaines conditions qui sont associées (en dur dans le code du logiciel) à la création des entités de classe Œuvre. Ces conditions impliquent par exemple la présence d'un titre uniforme dans chaque notice pour créer cette entité. Si ce titre uniforme n'est pas présent, aucune Œuvre n'est créée pour une notice donnée et donc toutes les autres entités associées ne sont pas migrées. Malheureusement, certains catalogues de notices ne contiennent que très

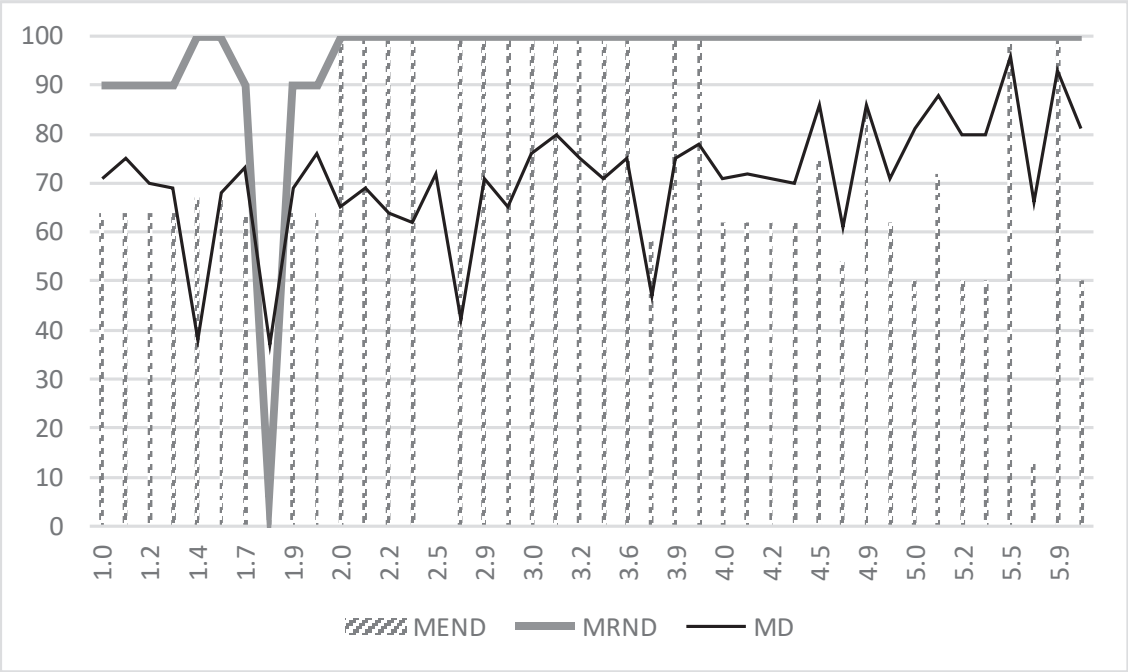


FIGURE 4.9 – Détection des motifs de connaissances par VFRBR

peu de titres uniformes [107], rendant ainsi les règles de VFRBR inutilisables. Concernant les motifs bibliographiques, l'orientation de l'outil sur le domaine musical lui impose des conditions spécifiques dans les règles de l'outil qui ne sont pas compatibles avec les motifs bibliographiques courants. En conséquence, l'outil ne permet pas de détecter correctement les motifs concernés par le jeu de données T42. De plus, le fait que les règles soient directement codées dans l'outil rend difficile leur réutilisation et leur adaptation dans d'autres contextes bibliographiques. Au final, cet outil est difficilement réutilisable en dehors de son contexte d'application initial.

La figure 4.10 présente les résultats de cette même expérimentation pour l'outil XC. Ici, la capacité d'interprétation des motifs par XC dépend fortement du motif évalué. Le motif basique (1.x) et les relations complémentaires (5.x) sont, pour la plupart des notices, correctement détectés. Pour le motif basique, la métrique MD (courbe grise) indique qu'environ 40% des données ne sont pas migrées. Cependant, les taux d'erreurs des métriques liées aux motifs (MEND et MRND) sont très bas, ce qui signifie que les données "oubliées" n'ont pas d'importance majeure pour la création des entités principales de FRBR, à la différence de VFRBR. En effet, XC crée systématiquement une entité de classe Œuvre, une entité de classe Expression et une Entité de classe Manifestation pour chaque notice traitée. Cela constitue la base de certains motifs, ce qui explique l'amélioration du score vis à vis de VFRBR. Concernant les problèmes de XC, les motifs de dérivations (3.x) et d'agrégations (4.x) ne sont pas correctement interprétés, essentiellement au niveau des relations principales. Les entités principales des motifs d'augmentations (2.x) ne sont jamais détectées mais les relations correspondantes peuvent être détectées mais associées à des entités non conformes avec l'expertise de T42. Ces résultats impliquent d'analyser plus en détail la méthode d'interprétation des augmentations dans les règles d'XC.

La figure 4.11 présente les résultats de l'expérimentation avec l'outil FRBR-ML. Ici les résultats sont similaires à ceux de XC. Le motif basique (1.x) est correctement détecté et modélisé avec un taux de données manquantes proche de 30%, ce qui est inférieur à celui d'XC. Le problème majeur réside dans la détection des relations principales des motifs avancées (MRND) qui est très souvent proche de 100%. Si les entités principales des motifs sont globalement détectées,

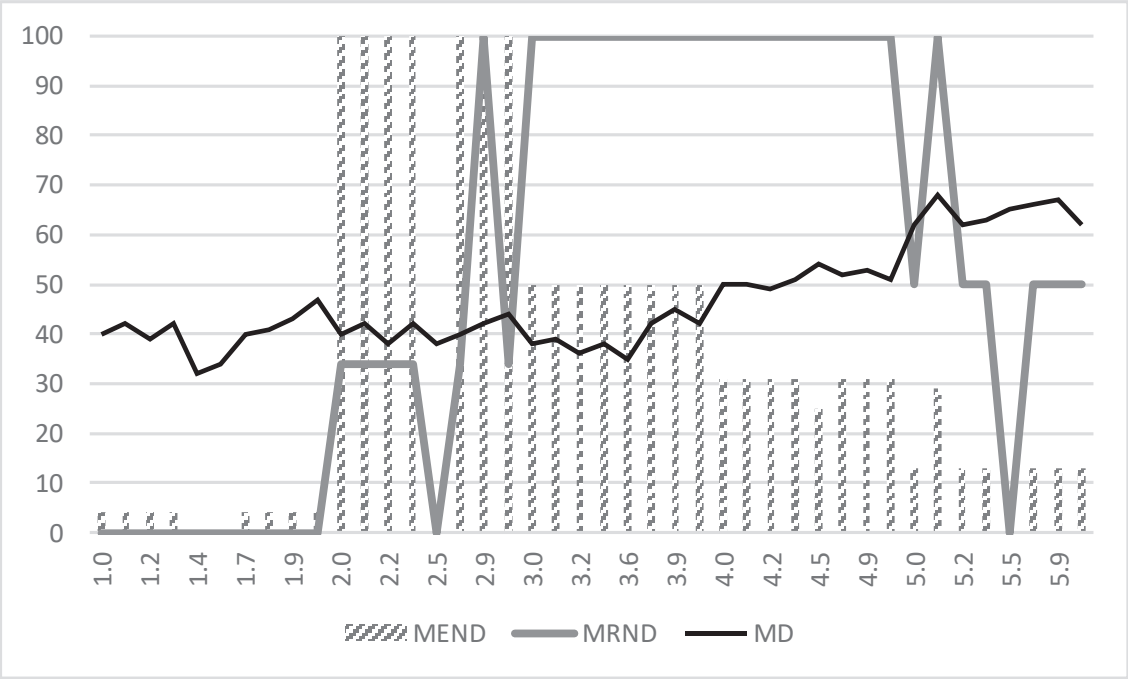


FIGURE 4.10 – Détection des motifs de connaissances par XC

à l'exception des augmentations où le motif n'est pas du tout détecté, l'outil ne créé pas les relations associées à ces entités. Nous pouvons toutefois noter que FRBR-ML détecte au moins la moitié des entités principales pour les motifs de dérivation et d'agrégation et obtient de bons résultats globaux sur la détection des éléments secondaires, mis à part les dérivations.

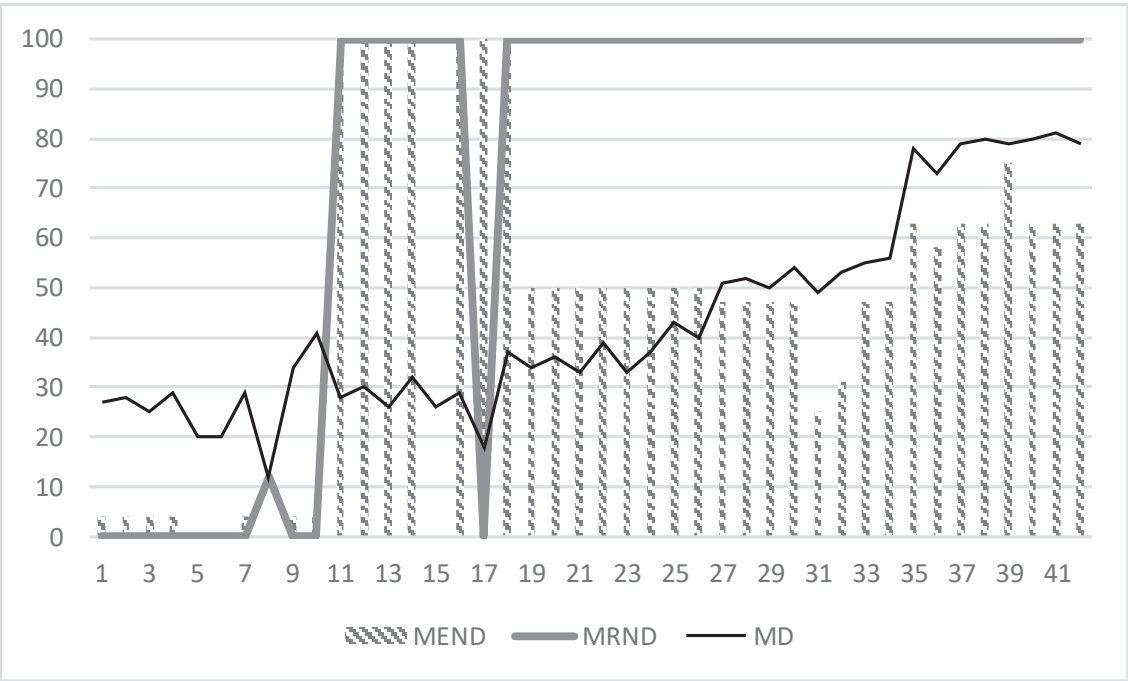


FIGURE 4.11 – Détection des motifs de connaissances par FRBR-ML

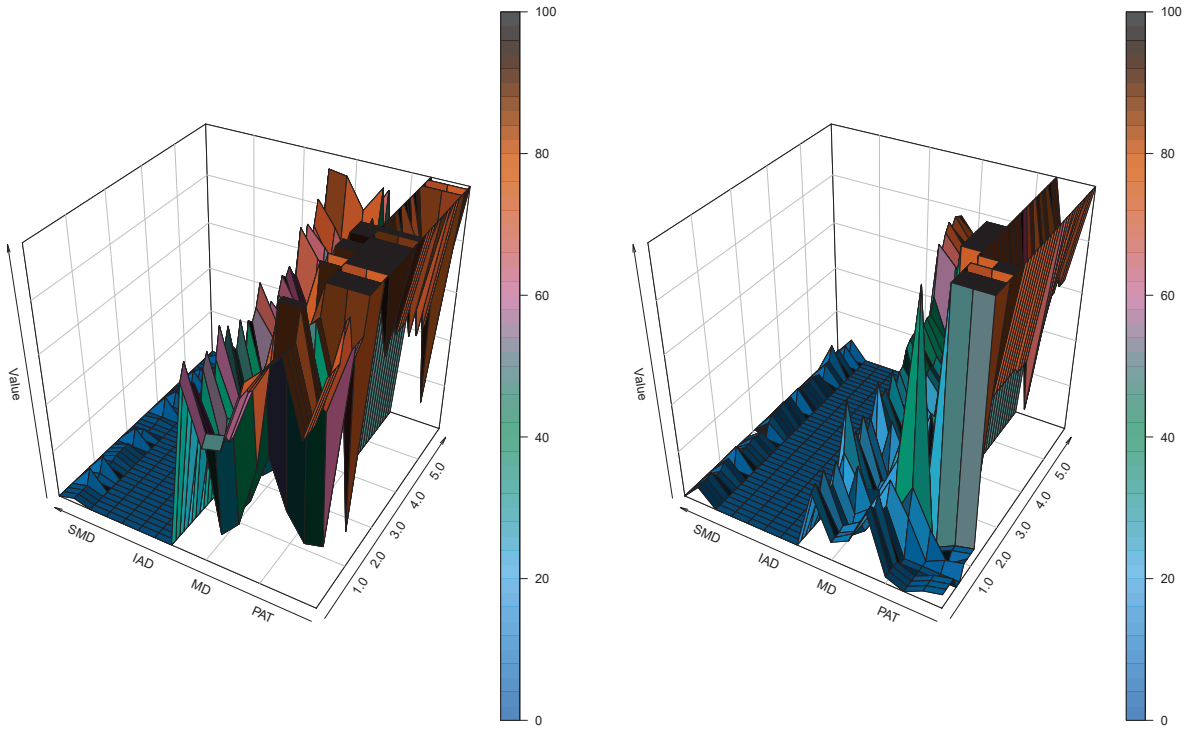
Le tableau 4.8 synthétise les éléments des motifs qui posent le plus de problèmes pour les trois outils. Par exemple, on observe le problème général des outils à identifier les relations biblio-

graphiques principales dans les motifs d'agrégation. Lorsqu'un outil combine des problèmes sur MEND et MRND pour le même motif, c'est que ce dernier n'est pas considéré par l'outil.

	Core	AUG	DER	AGG	COW
FRBR-ML	-	MEND, MRND	MRND, -	MRND	MRND
VFRBR	MRND, -	MEND, MRND	MEND, MRND, -	MRND	MEND, MRND
XC	-	MEND,MRND	MRND, -	MRND	MRND

TABLE 4.8 – Problèmes observés dans la détection des motifs de connaissances pour les 3 outils évalués

Les courbes présentés nous montrent que les outils de FRBRisation à disposition de la communauté ne prennent pas suffisamment en compte les spécifictés du domaine bibliographique comme les motifs de connaissances avancés. En conséquence, les modèles de règles de base qui sont proposés avec ces outils ne peuvent pas être utilisées pour réaliser un projet de FRBRisation avec une qualité satisfaisante. Nous distinguons cependant le potentiel de la solution FRBR-ML qui obtient des résultats meilleurs que les autres solutions et qui permet de facilement modifier les règles utilisées par l'outil au contraire de VFRBR et XC. La figure 4.12 confirme cette observation en représentant l'ensemble des résultats de T42 des outils, dans un diagramme en 3D, des outils VFRBR et FRBR-ML. L'analyse des couleurs dominantes donne une intuition sur les outils comme le fait ici que VFRBR tende plus vers des couleurs rouges et marrons (= plus mauvais résultats) et FRBR-ML tende plutôt vers du bleu (meilleurs résultats).



(a) Résultats de VFRBR sur T42 (b) Résultats de FRBR-ML sur T42

FIGURE 4.12 – Diagrammes en 3D de résultats d'expérimentations avec T42

De l'importance des métriques sur l'interprétation des connaissances

A partir des observations réalisées dans l'expérimentation précédente, nous réalisons une seconde expérience avec le jeu de données BIB-RCAT. L'objectif de cette expérience est d'observer les possibilités amenées par les métriques d'analyse d'un catalogue, que nous avons présenté dans ce chapitre, pour améliorer les règles de migration. Pour cela, nous évaluons le processus de migration des trois outils VFRBR, XC et FRBR-ML mais nous ajoutons une étape supplémentaire pour l'outil FRBR-ML.

Nous utilisons les métriques d'interprétation des métadonnées bibliographiques que nous appliquons au jeu de données BIB-RCAT pour identifier préalablement les spécificités des notices. Le calcul de ces métriques nous permet d'isoler la manière dont les motifs et anomalies sont présents dans ces notices et de discuter de la manière dont les connaissances peuvent être ainsi extraites et modélisées correctement dans les règles. Par exemple, nous avons calculé que 37% des notices de BIB-RCAT contiennent des pratiques spécifiques de catalogage. De la même manière, nous avons observé que 40% des notices contiennent des motifs d'augmentions des œuvres. Ensuite, à partir de ces statistiques, nous réalisons un nouveau jeu de règles pour l'outil FRBR-ML que nous appelons alors *FRBR-ML corrigé*. Le processus de correction consiste a adapter les règles initiales selon les possibilités de l'outil et les statistiques fournies dans un temps arbitraire de 4h (dont 3h pour modifier les règles et 1h pour réaliser des tests fonctionnels). Le résultat de cette expérimentation est présenté sur la figure 4.13.

Métrique	FRBR-ML	VFRBR	XC	FRBR-ML corrigé
MEND	94%	98%	94%	1%
MRND	100%	100%	100%	29%
ESE	99%	55%	100%	21%
MD	44%	45%	45%	13%

TABLE 4.9 – Résultats de FRBR-ML, VFRBR et XC sur le jeu de données BIB-RCAT

Dans ce tableau, nous observons des résultats similaires à ceux obtenus par les 3 outils avec le jeu de données T42. La moyenne de détection des éléments des motifs de connaissances (entité principale, relation principale et éléments secondaires) est proche de 100% d'erreurs avec les jeux de règles initiaux des outils. De plus, la quantité d'informations "oubliées" (MD) reste proche de 40%, comme à l'expérimentation précédente. L'aspect intéressant de ces statistiques réside dans les bons résultats obtenus par la version "corrigée" de FRBR-ML. En effet, en seulement 4h de temps, et grâce à l'analyse du catalogue, le taux de données oubliées passe de 44% à 13% et le taux d'entités principales des motifs passe de 94% d'erreurs à seulement 1% d'erreurs. Le diagramme de la figure 4.13 présente d'autres métriques de cettes expériences calculée entre FRBR-ML et FRBR-ML corrigé. Nous pouvons alors constater que l'adaptation des règles permet de corriger certains problèmes majeurs. Par exemple, la migration des relations de dérivation passe de 100% (pire cas) à 0% (cas parfait). Cela veut dire que l'outil a été capable de migrer correctement toutes les relations liées aux dérivations des œuvres. Ce progrès s'observe aussi par la métrique MR qui concerne les règles manquantes pour parfaitement migrer le catalogue. Le jeu de règle de FRBR-ML est passé de 24% de règles manquantes à seulement 7%.

Les résultats de cette expérimentation montrent l'intérêt des métriques d'interprétation des connaissances, que nous proposons dans ce chapitre, pour améliorer la qualité des processus de migration de notices bibliographiques. Les autres résultats de ces expérimentations sont répertoriés dans un document qui est disponible librement en ligne [8].

8. `http://bib-r.github.io/experiments.pdf`

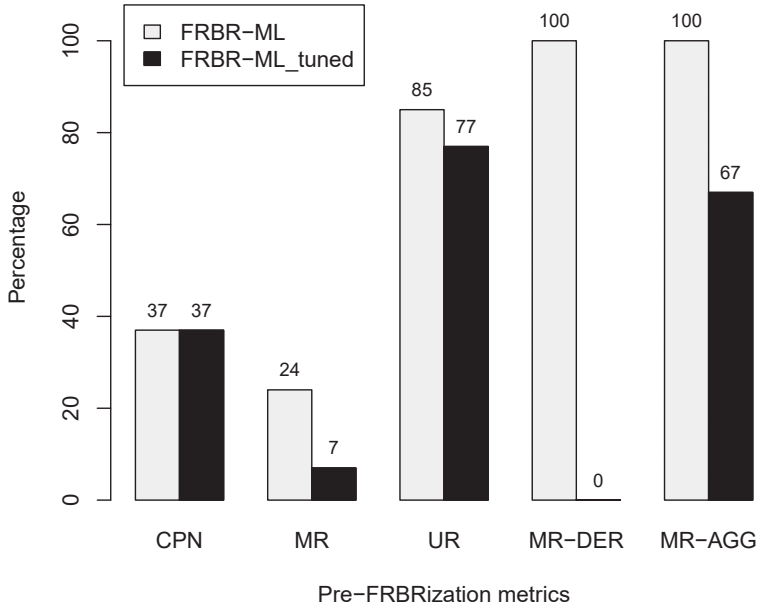


FIGURE 4.13 – Comparaison des résultats de FRBR-ML sur BIB-RCAT avec deux modèles de règles différents dont le second a été amélioré grâce aux métriques du chapitre 4.

4.6 Conclusion

Dans ce chapitre, nous avons présenté un ensemble de métriques couvrant le processus d'intégration de données bibliographiques. Nous avons également présenté comment nous avons intégré ces métriques dans un benchmark original, nommé BIB-R, permettant l'évaluation des solutions de FRBRisation. Les expérimentations réalisées avec ce benchmark ont mis en évidence les avantages de ces métriques pour l'amélioration de la qualité des outils de migration existants. La réflexion autour de cette perspective, dans la suite de ce manuscrit, porte sur une industrialisation de l'usage de ces métriques (et plus généralement des spécificités bibliographiques) dans la réalisation de projets qualitatifs de migration et d'enrichissement de données documentaires.

Dans le contexte de l'amélioration de la qualité de l'outil FRBR-ML, la manipulation des règles de cet outil a nécessité une expertise technique, car le langage XSLT utilisé peut être complexe à manipuler pour un non-spécialiste. Dans notre deuxième expérimentation, l'adaptation des règles a été effectuée de manière ponctuelle. Dans la réalité, il peut être nécessaire d'intégrer régulièrement des notices issues de divers catalogues, demandant ainsi des efforts plus importants pour paramétrer les règles. Il peut-être donc nécessaire de mettre en place une méthodologie plus particulière et plus complexe, incluant des outils supplémentaires pour faciliter l'intervention du documentaliste dans la gestion de ces règles. Ce type de méthodologie est l'objet d'étude du prochain chapitre dans ce manuscrit.

Chapitre 5

Méta-modélisation des connaissances

Ce chapitre présente une méthodologie originale pour extraire et modéliser les connaissances implicites d'un catalogue de notices bibliographiques, afin de permettre la migration automatique et l'enrichissement sémantique de ce dernier. Cette méthodologie repose sur l'utilisation d'un méta-modèle de migration qui tient compte des relations bibliographiques avancées qui sont implicitement cataloguées dans les notices documentaires.

Sommaire

5.1 Introduction

La transformation des catalogues bibliographiques issus des anciens formats hiérarchiques (notices) vers de nouvelles bases de connaissances sémantiques permet d'améliorer la réutilisation et la valorisation des ressources documentaires [107]. Lorsque les catalogues contiennent de nombreuses notices, cette transformation doit être réalisée de manière automatique en utilisant des règles permettant d'interpréter, de transformer et d'enrichir les métadonnées initiales. La figure 5.1 schématise l'intégration de ces règles comme des correspondances (réunies dans un modèle de migration) entre un modèle source (de type MARC) et le modèle cible (basé sur FRBR). Cette vision simpliste des règles ne peut cependant pas être appliquée telle qu'elle dans le domaine bibliographique. En effet, nous avons vu dans le chapitre 2 que la modélisation des relations

bibliographiques avancées du domaine impliquait des mécanismes particuliers [140, 5]. Ces mécanismes doivent permettre une interprétation des connaissances, décrites implicitement dans les notices, pour identifier les familles bibliographiques des entités d'un catalogue comme par exemple les Œuvres bibliographiques pouvant être agrégées, dérivées ou encore augmentées [79].

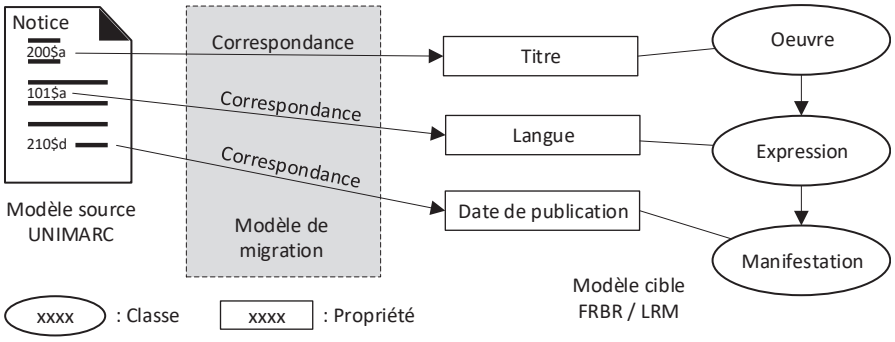


FIGURE 5.1 – Correspondances entre un modèle source UNIMARC et un modèle FRBR

Nous avons identifié au chapitre 3 que les outils existants de transformation des notices proposent des mécanismes avancés permettant de représenter des associations complexes entre un modèle source de notices et un modèle basé sur FRBR. Cependant, la prise en main de ces solutions, par les documentalistes reste complexe car certains outils ne permettent pas de modifier les règles de migration, d'autres reposent sur des formats de règles trop techniques ou trop limités et d'autres encore ont été conçus de manière générique et n'ont pas encore été appréhendés ni évalués dans le domaine bibliographique. L'utilisation de systèmes de migration non adaptés au contexte bibliographique peut entraîner la création de modèles de règles trop complexes et peu lisibles ayant des conséquences sur la qualité et la réussite du projet de transformation des métadonnées.

Pour faciliter la collaboration entre les experts informatiques et les experts du domaine, et ainsi simplifier la réalisation de projets de migration et d'enrichissement des catalogues, notre hypothèse est qu'il est nécessaire d'établir une interface pour les règles à un plus haut niveau d'abstraction. Cette interface doit permettre, d'une part, de visualiser les motifs bibliographiques qui doivent être intégrés dans la future base de connaissances par les documentalistes et d'autre part, d'instancier cette interface dans un outil de migration et d'enrichissement par des informaticiens. Il ne s'agit donc pas d'ajouter une couche de complexité à la modélisation des règles mais bien d'améliorer la lisibilité du modèle en encapsulant les correspondances de migration et d'enrichissement à un niveau d'abstraction des connaissances bibliographiques qui est connu et maitrisé par les documentalistes. L'objectif de cette contribution est ainsi de favoriser la réutilisation de règles dans la communauté en s'inspirant des standards bibliographiques et de proposer de nouveaux mécanismes pour faciliter la modélisation de cas spécifiques. En somme, cette contribution doit permettre une réconciliation des experts documentaires et des informaticiens dans leur approche technique de la migration et de l'enrichissement de métadonnées culturelles.

Prérequis sur la modélisation des connaissances

La gestion des connaissances, dans un domaine comme la documentation, s'opère à trois niveaux, le niveau *instance*, le niveau *modèle* et le niveau *méta*. En logique de description, ces niveaux sont respectivement appelés *ABox*, *TBox* et *MBox* [12]. ABox et TBox faisant référence respectivement aux données d'une base et à son schéma, nous développons l'hypothèse principale de ce chapitre sur le niveau MBox (correspondant au niveau d'abstraction à obtenir). Ce dernier consiste en une description des motifs de connaissances d'une base de données bibliographiques.

Nous prenons deux exemples de motifs de connaissances, l'*adaptation* et la *traduction*, que nous illustrons au niveau MBox dans la figure 5.2 à partir de l'œuvre *The Lord of the Rings*. Nous faisons remarquer que, pour l'ensemble des exemples de ce chapitre, nous utilisons la notion d'*entité* pour parler des objets du monde réel (ABox) (ex., *The Lord of the Rings*), la notion de *classe* pour parler des classes du modèle cible (TBox) (ex., Œuvre FRBR) et la notion de motif pour désigner les connaissances bibliographiques à intégrer (MBox) (ex., *Adaptation, Traduction*).

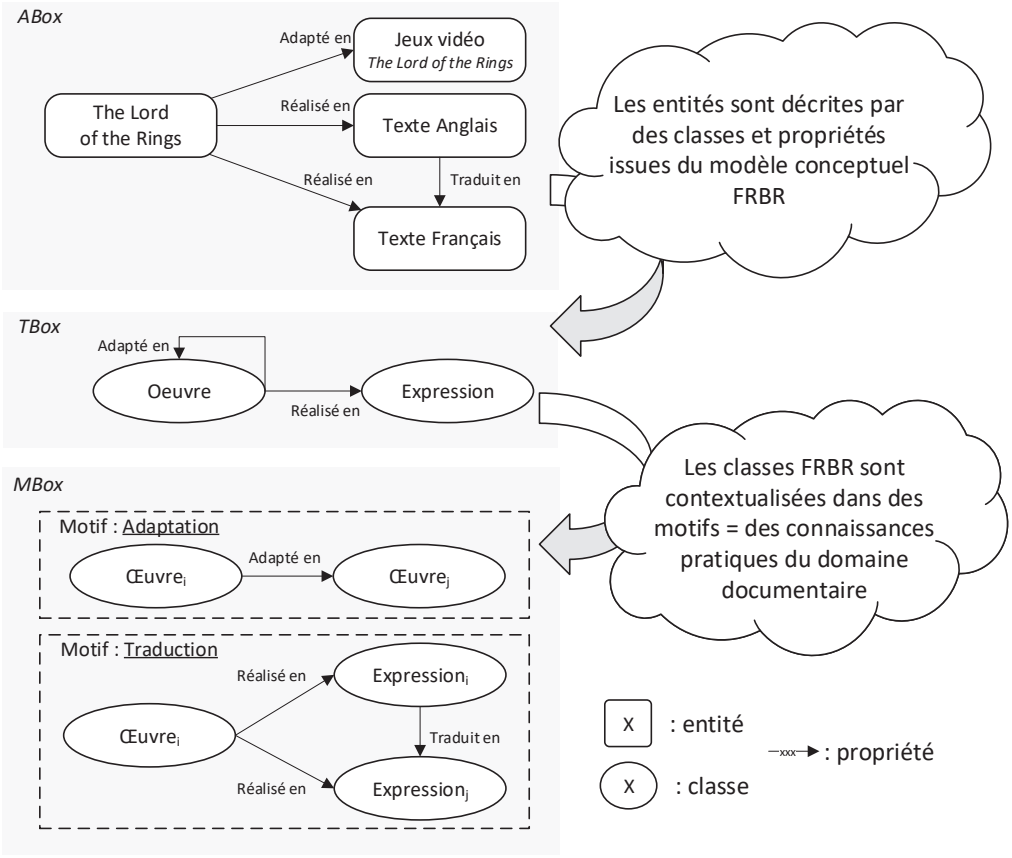


FIGURE 5.2 – Exemple appliqué aux trois niveaux de gestion de données

L'illustration de la figure 5.2 nous montre que les classes du modèle FRBR doivent être contextualisées pour révéler certains motifs de connaissances bibliographiques. Par exemple, une *Adaptation* peut être établie entre deux Œuvres distinctes dans un catalogue, l'Œuvre originale et son adaptation. De plus, une même classe peut être réutilisée dans différents motifs, comme ici l'Œuvre qui est utilisée à la fois dans le motif d'*Adaptation* et dans le motif de *Traduction*. Notre hypothèse, appliquée à cet exemple, défend que l'implication des documentalistes dans l'écriture des règles de migration et d'enrichissement nécessite une modélisation de ces dernières au niveau des motifs de connaissances. Or, cela demande un travail de contextualisation des connaissances entre le niveau TBox et le niveau MBox. Nous argumentons ci-après l'importance de ce travail de modélisation avant de présenter les objectifs de notre contribution dans ce contexte.

Les approches de Tzompanaki *et al.* et Britell *et al.* montrent que la méta-modélisation de données vers des méta-concepts, plus proches des connaissances des utilisateurs, facilite la création de règles pour l'intégration de connaissances depuis des sources de données externes [21, 135]. Dans le contexte des bases de données documentaires, qui peuvent être distribuées et hétérogènes, des projets d'intégration de données comme DELOS [7] ou EEXCESS [97] ont également recours à des modèles et ontologies conceptuelles (MBox) pour représenter les motifs de connaissances communs ayant un intérêt pour les utilisateurs. Cette modélisation (MBox) à un niveau

d'abstraction supérieur au modèle (TBox) permet également de faciliter l'interopérabilité entre les bases de données. Par exemple, différents travaux ont été menés pour la modélisation de motifs de connaissances culturels comme dans le domaine des musées [4, 72] ou plus généralement dans le web de données [96]. Dans le contexte des métadonnées bibliographiques, l'émergence des modèles de données conceptuels et sémantiques comme FRBR/LRM [114] offre des possibilités intéressantes pour valoriser les familles bibliographiques, qui sont considérées dans cette communauté comme la référence des motifs de connaissances des données bibliographiques [88].

Notre hypothèse s'appuie sur les problèmes soulevés au chapitre 4, concernant un manque de prise en compte des motifs bibliographiques dans les règles des outils de migration existants, et sur les avantages (cités plus haut) de la méta-modélisation des connaissances observés dans la communauté documentaire. La modélisation des règles au niveau des motifs de connaissances peut avoir un impact positif sur la faisabilité et la qualité des projets de transformation de catalogues bibliographiques. C'est pourquoi, dans ce chapitre, nous proposons une méthodologie originale pour la modélisation de ces règles qui répond à deux objectifs. Le premier objectif de notre méthodologie est de faciliter la prise en compte des motifs bibliographiques avancés afin d'améliorer la qualité des futures bases de connaissances dans le domaine documentaire. Notre second objectif est de réduire les efforts pour le paramétrage d'un processus de transformation automatique de catalogues documentaires en réutilisant des motifs pré-établis. Notre méthodologie intègre des mécanismes de modélisation qui sont essentiels pour la réalisation de ces objectifs.

Dans le reste de ce chapitre, la section 5.2 détaille les caractéristiques de notre modélisation de règles pour la migration de notices bibliographiques. La section 5.3 présente une extension de notre modèle de règles pour l'enrichissement des entités migrées avec des sources de données externes. La section 5.4 décrit l'application du modèle de migration et d'enrichissement pour des motif fréquents et des cas plus spécifiques de connaissances bibliographiques issues de catalogues documentaires. Enfin, la section 5.6 discute des avantages et perspectives de notre contribution.

5.2 Conception d'un modèle de migration

Dans cette section, nous décrivons notre méthodologie pour créer un modèle de migration qui permette la transformation d'un catalogue de notices en respectant les enjeux du domaine bibliographique. Ce modèle de migration repose sur une interprétation des connaissances avancées d'un catalogue (ex., adaptations, traductions, illustrations) pour encapsuler les règles de migration dans un graphe orienté de motifs de connaissances.

5.2.1 Vue d'ensemble

Notre approche consiste à associer les données d'un catalogue bibliographique avec un ensemble de motifs de connaissances correspondant aux familles bibliographiques à représenter dans la base de connaissances. Ces connaissances sont supposées connues et organisées par la communauté bibliographique [137, 79, 133]. Pour le reste de ce manuscrit, nous appelons ce réservoir de motifs, le *méta-modèle des connaissances bibliographiques*.

Définition 5.2.1 (Méta-modèle des connaissances bibliographiques)**.** Ensemble de motifs conceptuels de connaissances bibliographiques correspondant aux différentes informations ayant un intérêt pour un utilisateur d'un catalogue documentaire.

Prenons en exemple une notice de l'œuvre de *Tolkien*, *The Lord of the Rings*, traduite en français. La traduction de cette œuvre a un intérêt pour l'utilisateur et doit être clairement représentée dans la base de connaissances. La Figure 5.3 illustre notre approche en présentant la migration de

cette notice, au format UNIMARC, en utilisant trois motifs, respectivement *Création*, *Traduction* et *Traducteur*, issus du méta-modèle des connaissances. Ces motifs encapsulent les règles qui permettent d'organiser les connaissances telles qu'elles doivent être représentées dans la future base de connaissances. Ainsi, la création du modèle de règles pour la migration, c'est à dire une instance du méta-modèle des connaissances, repose sur deux étapes, l'*interprétation* des champs du modèle source et la *modélisation* des motifs avec les éléments du modèle cible.

Exemple 5.2.1. Dans l'exemple de la Figure 5.3, un des motifs de connaissance est la *Création*. L'étape d'interprétation de la notice consiste ici à associer les champs 454$t (titre original), 700$a (nom de la responsabilité) et 700$4 (fonction de la responsabilité), dont le code "070" signifie *auteur* pour obtenir les informations minimales sur la création de l'œuvre. L'étape de modélisation consiste à proposer une représentation du motif *Création* à partir des éléments du modèles cible de la migration. Ici nous faisons le choix d'utiliser une classe Œuvre, liée à une classe Agent de FRBR pour produire respectivement les entités *The Lord of the Rings* et *Tolkien*.

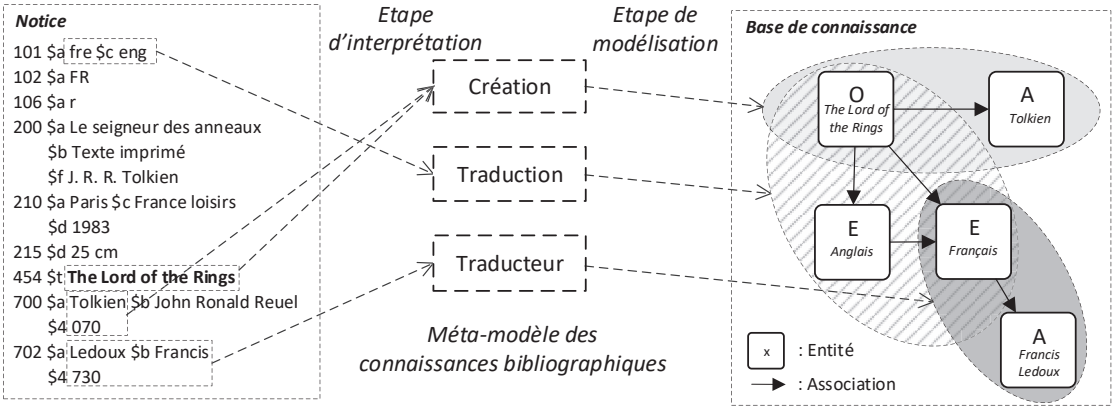


FIGURE 5.3 – Exemple d'usage des motifs de connaissances dans le processus de migration.

Nous présentons maintenant comment ce travail d'interprétation et de modélisation des connaissances (résultant en un méta-modèle comme décrit plus haut) peut être intégré dans le processus concret de migration des notices bibliographiques. La figure 5.4 illustre notre approche à l'échelle d'un catalogue bibliographique composé de plusieurs notices. En entrée du processus nous avons le catalogue et le méta-modèle pré-établi des connaissances bibliographiques. L'étape (1) d'analyse du catalogue permet de modéliser une instance du méta-modèle, l'étape (2) consiste à former un modèle de migration qui sera utilisé à l'étape (3) pour transformer les notices en une base de connaissances sémantiques.

Sur la Figure 5.4, la phase (1) consiste en l'analyse des spécificités du catalogue. Cette phase permet de déduire les motifs de connaissances qui sont pertinents et de définir des fonctions pour interpréter les notices. Par exemple, si les notices contiennent des informations sur des traductions et des adaptations d'œuvres, l'analyse permet de détecter les motifs *Création*, *Adaptation*, *Traduction* et *Traducteur* dans un méta-modèle prédéfini. Ce méta-modèle, basé sur les standards de la communauté (*ex.*, FRBR) définit la manière de structurer l'information pour chaque motif. La phase (2) consiste à créer une instance de ce méta-modèle, appelée le *modèle de migration*, pour définir l'organisation des motifs de connaissances afin d'automatiser la migration des notices. Cette phase définit les relations et conditions qui peuvent être appliquées entre les motifs ainsi que la manière dont les classes et propriétés du modèle cible sont agencées au sein

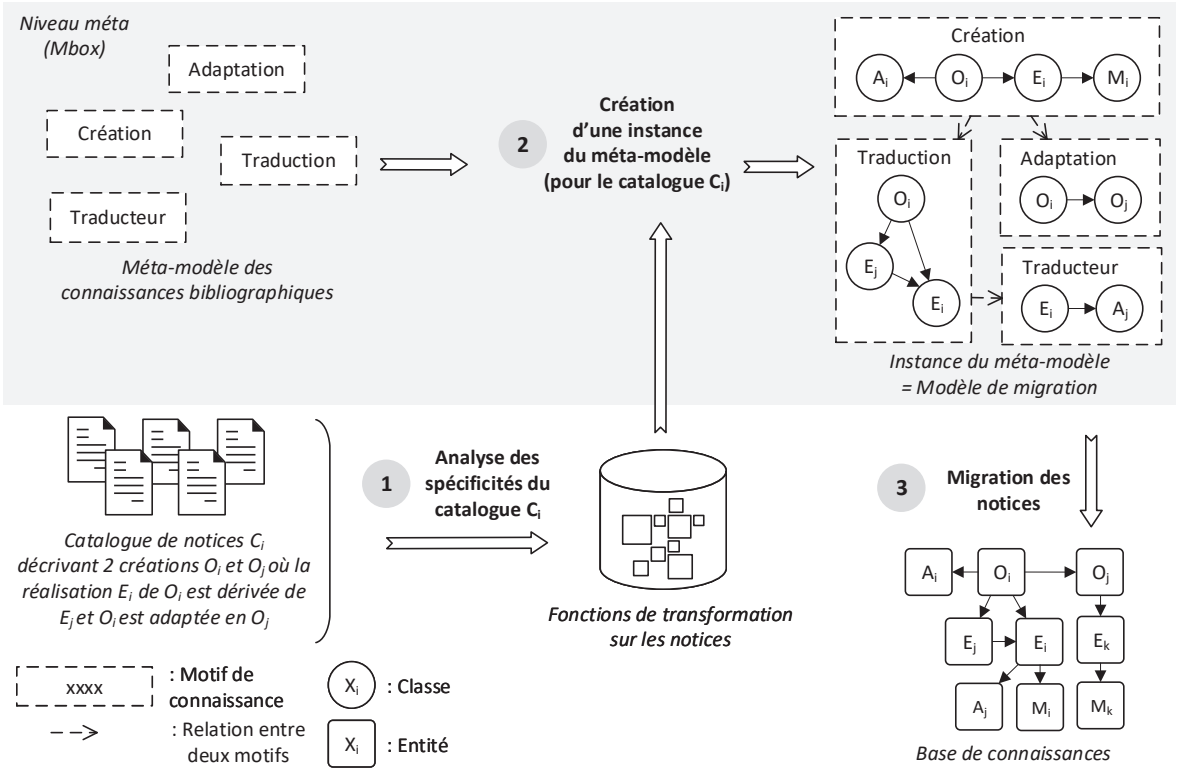


FIGURE 5.4 – Création et application d'un modèle de migration

de chaque motif. La dernière phase (3) consiste en l'implémentation de l'instance dans un outil de migration pour réaliser la création automatique de la base de connaissances.

Exemple 5.2.2. Une relation possible entre un motif de *Création* et un motif de *Traduction* peut être soumise à la condition que, pour une notice traitée par un outil de migration, il existe une zone de responsabilité secondaire (702 en UNIMARC) qui possède le code de fonction *traducteur* (730 en UNIMARC). La manière de construire et de représenter ces relations entre motifs est décrite plus tard dans cette section.

5.2.2 Description du modèle de migration

Dans cette partie, nous détaillons formellement les composants du modèle de migration, qui font correspondre les éléments du modèle source avec ceux du modèle cible. Dans notre contexte, le modèle source est un modèle composé de notices bibliographiques (détaillées dans le chapitre 2). Le modèle cible est une base de connaissances sémantiques composée de classes et de propriétés.

Concepts élémentaires

Nous commençons par définir les différents concepts essentiels qui interviennent dans la création d'un modèle de migration bibliographique classique.

Définition 5.2.2 (Contextes bibliographiques)**.** Les contextes bibliographiques d'une entité sont les différentes connaissances intellectuelles et éditoriales qui peuvent être liées à cette entité dans un catalogue bibliographique. Chaque entité possède donc plusieurs contextes dans un catalogue donné, ces derniers pouvant être implicitement représentés par différents champs dans plusieurs notices bibliographiques.

Exemple 5.2.3. Un contexte bibliographique peut s'incarner par des relations entre une Œuvre et ses différentes réalisations ou encore par des liens entre un Agent et d'autres entités d'un catalogue. Un contexte bibliographique pour l'Œuvre *The Lord of the Rings* de *J.R.R. Tolkien*, dans une bibliothèque française, peut inclure une édition traduite en français, ses adaptations en films, une relation avec une œuvre qui décrypte l'univers de l'auteur ou encore une adaptation en bande dessinée de l'œuvre originale. Dans l'exemple de la figure 5.3, la notice proposée décrit un contexte bibliographique l'Œuvre *The Lord of the Rings* qui est sa traduction en français.

L'ensemble des relations et des entités d'un même contexte peuvent être décrites au sein d'une même notice ou dans plusieurs notices. C'est pourquoi la modélisation de ces contextes fait intervenir à la fois des mécanismes d'interprétation d'une notices mais nécessitent également un processus complémentaire de déduplication. Dans ce chapitre, nous nous concentrons sur les mécanismes appliqués à l'échelle d'une notice à transformer.

Définition 5.2.3 (Notice)**.** Une notice est une unité d'information dans un catalogue bibliographique qui est constituée d'un ensemble de champs pour décrire les métadonnées d'une ressource possédée par une institution documentaire. Nous considérons une notice n comme un ensemble de couples (c_i, v_i) de clés et de valeurs. Les champs des notices peuvent être hiérarchises dans des zones répondant à une organisation éditoriale qui n'a pas d'influence sur l'interprétation des connaissances intellectuelles dans notre contexte.

Définition 5.2.4 (Fonctions d'interprétation)**.** Nous considérons l'ensemble F des fonctions d'interprétations permettant l'extraction des données pour leur intégration dans une base de connaissances. Chaque fonction d'interprétation $f \in F$ prend en entrée un ou plusieurs champs et retourne une valeur, associée à un label, qui peut être exploitée.

Exemple 5.2.4. Selon l'exemple de la figure 5.3, une fonction d'interprétation, permettant la détection d'une traduction, peut prendre en entrée les champs UNIMARC 454$t (titre original), 700$a (nom de la responsabilité) et 700$4 (fonction de la responsabilité) et retourner une valeur booléenne décrivant si une notice décrit une traduction ou non.

Définition 5.2.5 (Base de connaissances)**.** Une base de connaissances sémantiques est un graphe G où chaque nœud représente une classe et chaque arc représente une propriété caractérisant un lien entre deux classes. Une classe contient également des propriétés vers des littéraux. C'est pourquoi, nous considérons chaque classe du graphe G comme un ensemble de couples (p_i, σ_i) de propriétés p_i menant à une autre classe du graphe ou à un littéral, où l'URI de la classe ou bien le terme sont symbolisés par σ_i.

Pour des raisons de lisibilité, les prochaines définitions agrègent les éléments constitutifs d'un graphe modèle d'une base de connaissances, c'est à dire les notions de classe et de propriété, dans la notion abstraite d'*élément* d'un graphe G noté n. Par exemple, la classe Œuvre et la propriété *titre* de l'Œuvre sont des éléments composant le modèle conceptuel de FRBR (qui est un graphe) que nous incluons dans un ensemble $\{n_j\}_j$.

Composants du modèle de migration

Nous définissons désormais les différents composants du modèle de migration, ou dit autrement, la structure des règles permettant de migrer un catalogue de notices. Ces composants s'articulent autour des principes d'interprétation et de modélisation que nous avons présentés précédemment. Plus précisément, ils sont basés sur des correspondances, qui, à la différence des approches traditionnelles, associent des informations bibliographiques au niveau méta (TBox -> MBox) au lieu de lier des éléments au niveau du modèle (TBox -> TBox). Cela veut dire que les correspondances pour la migration associent une combinaison de plusieurs champs issus d'une notice, d'une part, avec une combinaison de plusieurs classes et propriétés du modèle cible, d'autre part.

Définition 5.2.6 (Correspondance de migration)**.** Une correspondance de migration ϕ est un triplet d'ensembles : $\phi =< \{c_i\}_i, \{n_j\}_j, \{f_k\}_k >$ où $\{c_i\}_i$ est un ensemble de clés d'une notice, $\{n_j\}_j$ un ensemble de classes ou propriétés et $\{f_k\}_k$ un ensemble de fonctions d'interprétations.

Exemple 5.2.5. Considérons un catalogue composé de plusieurs notices décrivant des œuvres et des agents qui sont des contributeurs des Œuvres. Les Agents sont décrits dans les notices par les champs 200$\$f$ et 200$\$g$. Un exemple d'instanciation pour ces champs peut s'écrire : 200$\$f$ = *"Bob Smith, John Rel"* et 200$\$g$ = *"Dan Browder, Jason Burth, Sam Gamgee"*. Nous proposons alors la correspondance de migration suivante :
$\phi_{contributeurs} =< \{200\$f, 200\$g\}, \{(contributor(uvre) \rightarrow Agent)\}, f_{noms}\}$.
Pour modéliser la relation entre l'Œuvre, décrite dans une notice, et les contributeurs, nous utilisons les classes issues de FRBR, *Oeuvre* et *Agent*. Nous ajoutons ensuite la relation de contribution en utilisant la propriété nommée *contributor*. De plus, les valeurs des deux champs contenant plusieurs noms d'entités séparés par des virgules, nous utilisons une fonction d'interprétation f_{noms} qui effectue une opération de type *split*(","), afin d'isoler les différents noms, créer un Agent pour chaque nom et associer chaque Agent à l'Œuvre par la relation *contributor*. Nous notons que la manière dont une fonction f doit utiliser les données $\{c_i\}_i$ est implicite ici mais relève principalement de spécificités d'implémentation des fonctions $\{f_k\}_k$ dans l'outil de transformation utilisé.

Dans notre approche, les correspondances de migration, basées sur des éléments issus du modèle source et du modèle cible (TBox), sont encapsulées pour former des motifs décrivant un contexte bibliographique, à un plus haut niveau d'abstraction (MBox). Cette modélisation permet de représenter l'ensemble des connaissances intellectuelles que l'utilisateur souhaite exploiter en conservant la notion simple de *correspondance*.

Définition 5.2.7 (Motif de connaissances du modèle de migration)**.** Un motif de connaissances, dans l'instance du modèle de migration, $m =< \sigma_i, \{\phi_i\}_i >$ est défini comme l'association d'un label σ_i et d'un ensemble de correspondances $\{\phi_i\}_i$.

Nous précisons ici que cette définition s'applique dans la version "instanciée" du modèle de migration car, dans le domaine bibliographique théorique, un motif de connaissances peut également désigner une simple modélisation d'une connaissance sans forcément intégrer la notion de correspondance. Dis autrement, un *motif de connaissances du modèle de migration*, c-à-d dans notre méthodologie, fait parti intégrante des règles de migration et est utilisé concrètement par l'outil pour transformer les notices.

Exemple 5.2.6. La Figure 5.5 illustre la notion de motif de connaissances avec un exemple de motif σ_k = *Traduction*, qui est créé à partir de la combinaison de deux correspondances ϕ_i et ϕ_j, décrivant respectivement la relation de *réalisation* d'une œuvre et de *traduction* d'une réalisation.

Dans l'exemple de la figure 5.5, nous montrons comment un motif de migration, dans notre approche, peut se construire à partir de correspondances (pour conserver une compatibilité avec les outils existants de migration). Dans cet exemple, la première correspondance peut s'écrire : $\phi_i =< \{c_i\}, \{(réalisation(O_i) \rightarrow E_i)\}, f_{typeFormat}\}$ où c_i est un champ, issu du modèle des notices, décrivant le type de format pour la publication d'une œuvre. O_i, *réalisation* et E_i sont issus du modèles cible et $f_{typeFormat}$ est une fonction d'interprétation pour décrypter le type de format issu de la valeur v_i associée à c_i. La seconde correspondance peut s'écrire $\phi_j =< \{c_j\}, \{(traduction(E_j) \rightarrow E_i)\}, f_{langue}\}$ où c_j est un champ contenant une langue de traduction. Les Expressions E_j, *traduction* et E_i sont issus du modèle cible et $f_{traduction}$ est une fonction d'interprétation de la langue exposée par la valeur v_j associée à c_j. La fusion de E_i dans les deux correspondances ϕ_i et ϕ_j et l'application de ϕ_i avec E_j permet de former le motif de connaissances σ_k = *Traduction*. Les mécanismes permettant cette fusion sont détaillés plus loin

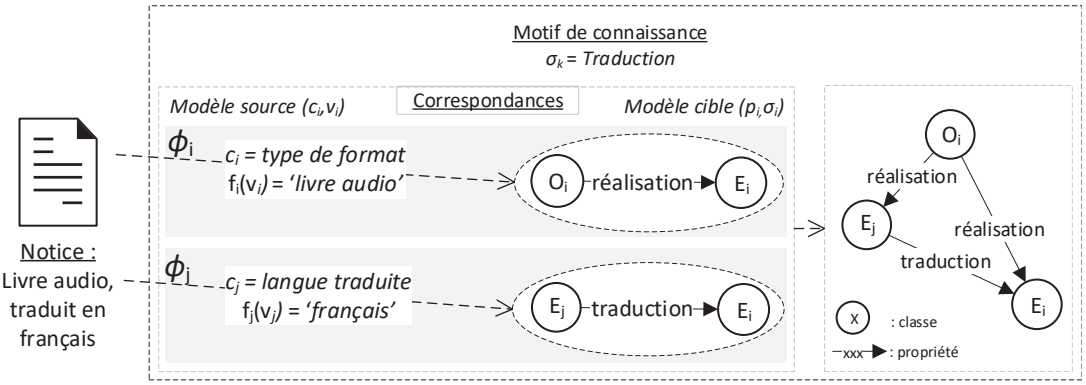


FIGURE 5.5 – Construction d'un motif de connaissances à partir de correspondances de migration

dans ce chapitre. Le motif *Traduction*, nouvellement créé, peut être ensuite documenté et réutilisé pour la migration de différents catalogues de notices bibliographiques.

Nous avons défini la notion de motif de connaissances, permettant l'encapsulation des correspondances de migration. Ces dernières contiennent les informations permettant d'interpréter les notices et chaque motif, contenant un ensemble de correspondances, permet d'exprimer une connaissance qu'il convient de modéliser à l'issue du processus de migration. L'ensemble des motifs forme le méta-modèle des connaissances bibliographiques.

Définition du modèle pour la migration

Nous définissons maintenant le modèle de migration, c'est à dire une *instance spécifique du méta-modèle* (de motifs) permettant la migration automatique d'un catalogue bibliographique. Cette instance doit répondre à deux principes qui sont respectivement (1) les objectifs métier de la future base de connaissances et (2) la prise en compte des spécificités du catalogue à migrer. Pour le premier principe, il s'agit d'articuler les connaissances autour d'une *entité primaire* (ou encore *progenitor*, *cf.* Chapitre 2) à partir de laquelle les utilisateurs peuvent explorer les connaissances du catalogue. Dans le second cas, les motifs de connaissances sont organisés, c'est à dire qu'ils sont assemblés et ordonnés en un arbre dont la racine contient la description de l'entité primaire, les nœuds sont les motifs de connaissances sélectionnés et les arcs représentent les conditions, spécifiques au catalogue à migrer, qui activent les motifs pertinents pour une notice donnée.

Définition 5.2.8 (Entité primaire)**.** Une entité primaire est définie par une classe issue du modèle cible autour de laquelle s'articulent les motifs de connaissances.

Exemple 5.2.7. Dans un catalogue bibliographique, les connaissances peuvent s'articuler autour de concepts élémentaires comme les Œuvres ou les Agents du domaine bibliographique. Par exemple, dans l'objectif de représenter des familles bibliographiques, en termes de dérivations et d'agrégations de ressources documentaires, l'entité primaire peut être basée sur la classe Œuvre telle que nous l'avons définie dans le chapitre 2.

Définition 5.2.9 (Instance du méta-modèle)**.** Une instance du méta-modèle des connaissances bibliographiques est un arbre enraciné $A =< E_{prim}, C, M >$ dont les nœuds M sont l'ensemble des motifs de connaissances sélectionnés pour la migration, les arcs sont issus d'un ensemble de conditions C permettant l'activation des motifs pertinents, lors du traitement d'un catalogue, et E_{prim} est l'entité primaire dont la classe, dans le modèle cible, est décrite dans le nœud racine de l'arbre.

La figure 5.6 illustre la notion d'instance de méta-modèle en montrant un exemple d'arbre de motifs. Dans cet exemple, la racine est un motif dont le label σ_i est *Œuvre Bibliographique* et qui permet de modéliser une Œuvre FRBR et ses attributs principaux. A partir de ce motif *Œuvre Bibliographique*, deux conditions d'interprétation c_j et c_m sont symbolisées par deux arcs qui mènent respectivement aux motifs *Réalisation de l'Œuvre* et *Sujets de l'Œuvre*. Lors de l'interprétation d'une notice bibliographique, si les deux conditions c_j et c_m sont vérifiées, les deux motifs nommés σ_j et σ_m seront évalués, c'est à dire que les correspondances qu'ils encapsulent seront appliquées.

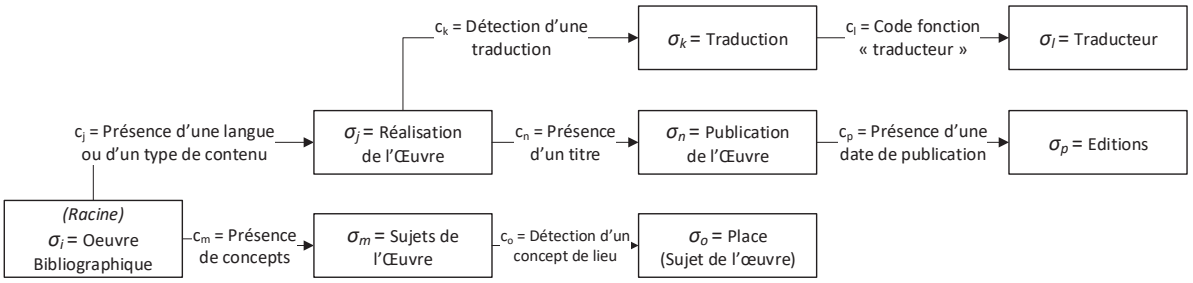


FIGURE 5.6 – Exemple du méta-modèle instancié sous la forme d'un arbre de motifs

Nous rappelons, que le modèle de migration, c'est à dire l'instance du méta-modèle, décrit et structure l'ensemble des connaissances qui doivent être extraites des notices d'un catalogue qui est migré. Comme les notices d'un catalogue décrivent chacune des ressources différentes, leur évaluation selon le modèle de migration va engendrer un parcours spécifique (pour la notice) dans l'arbre des motifs durant la migration. Dis autrement, toutes les notices évaluées n'activent pas nécessairement l'ensemble des motifs du modèle, mais seulement ceux qui concernent cette notice selon les conditions définies dans l'arbre. Les motifs pouvant être réutilisables entre différents catalogues, nous préconisons d'analyser en amont un catalogue à migrer (en utilisant les métriques décrites au chapitre 4) afin de limiter le modèle aux seuls motifs qui sont utiles pour ce catalogue. Cela permet d'améliorer la lisibilité globale du modèle pour les documentalistes et de réduire le nombre de conditions à évaluer lors du traitement du catalogue (i.e., éviter les calculs inutiles de conditions non applicables pour un catalogue).

Plusieurs instances du méta-modèle (donc plusieurs arbres) peuvent être crées pour réaliser la migration d'un catalogue. La particularité de chaque arbre va dépendre de l'entité primaire choisie, de l'agencement des motifs de connaissances et de la complexité de ces derniers. Dans l'exemple précédent, nous décidons d'utiliser la notion d'Œuvre comme racine de l'arbre mais cela pourrait être la notion de Manifestation. De plus, nous choisissons de séparer les motifs de traduction et de traducteur pour améliorer la lisibilité du modèle mais ils pourraient également être fusionnés. L'idée à retenir est que notre méta-modélisation est conçue de manière flexible pour offrir une représentation des règles qui soit cohérente pour les documentalistes. De plus, cette modélisation peut-être facilement implémentée dans des outils de migration comme X3ML ou Karma qui permettent de gérer des correspondances dont la partie droite peut-être un graphe de classes et propriétés (comme les correspondances des motifs de notre méthodologie) [83, 71].

5.3 Extension du méta-modèle pour l'enrichissement

Nous avons observé, dans le chapitre 3, que l'enrichissement des entités FRBR avec des sources externes était un processus essentiel pour améliorer la qualité des bases de connaissances bibliographiques et pour multiplier les liens entre ces bases et le reste du web de données. Pour rappel,

le processus d'enrichissement est envisagé de différentes manières selon la nature des données disponibles. Certaines informations utiles sont contenues dans des textes bruts dont il faut extraire les propriétés à exploiter [128, 136, 117]. D'autres connaissances sont obtenues par l'intégration de concepts issus du web de données [26, 57]. Enfin, d'autres formes d'enrichissements sont réalisées en classifiant les données existantes avec d'autres concepts issues d'ontologies ou de thésaurii [78, 63].

L'implémentation de notre modèle de migration, détaillé précédemment, permet la construction d'une base de connaissances contenant des métadonnées issues de la migration des notices. Ces métadonnées sont structurées selon les motifs de connaissances (*ex.*, *adaptions*, *agrégations* ou *traductions*) qui ont été définis dans le modèle de migration. Chaque motif décrit donc des relations sémantiques qu'il peut être intéressant de récupérer, pas uniquement des notices locales, mais de sources externes. Par exemple, le motif d'*adaptation* peut contenir la relation *adaptedAsWork*, du vocabulaire *RDA*, qui correspond à la propriété *BasedOn (P144)* de Wikidata. Dans ce cas, le motif de notre modèle peut décrire cette équivalence pour permettre un enrichissement des données locales avec Wikidata. Nous proposons donc de réaliser l'enrichissement des connaissances des métadonnées locales en réutilisant la modélisation existante des motifs de connaissances.

Plus particulièrement, nous proposons de spécialiser la notion de *correspondance* du modèle de migration et de créer des *correspondances d'enrichissement* dans les motifs du méta-modèle. Comme pour les notices bibliographiques, il peut y avoir une différence de modélisation entre les données locales (c-à-d, des entités et relations FRBR) à enrichir et les connaissances externes. D'ailleurs, dans le web de données, les relations bibliographiques avancées entre les entités (*ex.*, augmentations d'Œuvres) ne sont que peu représentées ou implicites, même si certaines bases de connaissances comme *data.BNF* commencent à diffuser ces informations de manière explicite. L'objectif des correspondances d'enrichissement est de permettre la génération de requêtes d'enrichissement pour l'intégration de nouvelles connaissances bibliographiques dans ce contexte.

5.3.1 Prérequis

Nous commençons par présenter les principes élémentaires dans la création d'un processus d'enrichissement. Ces principes influencent la manière dont le modèle d'enrichissement, c-à-d les règles pour l'extraction et l'intégration des données, doit être structuré.

Les informations locales représentent les données en entrée du processus d'enrichissement. Ces informations précisent le périmètre intellectuel de l'enrichissement ainsi que la manière d'intégrer les données. Dans notre cas, nous considérons que l'enrichissement s'effectue sur une seule entité locale à la fois (*ex.*, une Œuvre, un Agent). De plus, les données existantes sur cette entité sont supposées déjà structurées en entités et associations.

Les modèles de données distants spécifient la manière dont l'information à extraire est structurée dans les sources externes. Si, pour la migration (section précédente), les données à intégrer sont issues de notices bibliographiques, les modèles de données distants pour l'enrichissement, dans notre contexte, sont supposés structurés dans un graphe RDF.

Définition 5.3.1 (Graphe RDF)**.** Soit $T = I \cup L \cup B$ l'ensemble des vocabulaires d'une base de connaissances où I, L et B sont respectivement des ensembles d'IRIs, de Littéraux et de *Blank Nodes*[1]. Un graphe RDF est un ensemble fini de triplets RDF où chaque triplet (s, p, o) est défini sur $T \times I \times T$ avec s représentant un sujet, p, un prédicat et o, un objet.

1. `https://www.w3.org/TR/rdf11-concepts/#section-blank-nodes`

La **spécificité des requêtes d'enrichissement** influence la quantité et la pertinence des données qui sont extraites des sources avant d'être intégrées dans la base locale. Adapter cette spécificité revient à décider si la création des requêtes se fait dans une hypothèse dite *du monde clos* (*closed-world assumption*) ou du monde ouvert (*open-world assumption*). Dans le premier cas, les requêtes sont orientées pour n'extraire qu'une information ciblée par les motifs de connaissances, ces derniers définissant les limites des informations supposées valides. Dans le second cas, les requêtes permettent l'extraction d'une quantité plus large de données afin qu'un processus dédié réalise une validation *a posteriori* sur les données à intégrer. Nous privilégions l'hypothèse du monde clos dans le contexte d'une FRBRisation d'un catalogue de notices (c'est le contexte de ce chapitre) afin de contrôler les données intégrées aux entités produites. Dans un second temps, lorsque la nouvelle base de connaissances FRBR est en production, nous pouvons envisager la mise en place d'une fonctionnalité d'exploration de nouvelles connaissances en monde ouvert.

Pour résumer notre système initial, nous considérons une base de connaissances locale contenant des entités et des propriétés à enrichir (issues de la migration d'un catalogue). Chaque processus d'enrichissement est réalisé à partir d'une des entités de cette base que nous enrichissons par l'application d'un ensemble de requêtes sur différents graphes RDF externes. Ces requêtes sont construites à partir d'un modèle d'enrichissement qui est lui-même basé sur un méta-modèle des connaissances bibliographiques. Le procédé est similaire au modèle de migration de la section précédente. Nous le détaillons brièvement :

Un méta-modèle prédéfini utilise les classes et propriétés du modèle local de l'entité à enrichir (niveau TBox) pour décrire les connaissances locales (niveau MBox) que l'on souhaite enrichir à partir de sources distantes. Par exemple, nous avons migré une entité de classe Œuvre FRBR par la migration d'une notice (*cf.*, section précédente) et nous connaissons déjà le motif d'*Adaptation* des Œuvres (voir la figure 5.2). Nous n'avons pas d'instance d'adaptations dans le catalogue migré mais nous souhaitons détecter des instances dans des sources externes. Pour cela, nous devons associer la description des motifs locaux pour l'enrichissement (*ex.*, *Adaptation*) avec les modèles des sources distances (*ex.*, des graphes RDF). C'est le rôle du modèle d'enrichissement, c'est à dire l'instance du méta-modèle, de décrire ces associations. De manière très concrète, le modèle d'enrichissement peut, par exemple, décrire une association entre un motif d'*Adaptation* dans la base de connaissances locale, et le sous-graphe de classes et propriétés décrivant des adaptations dans des bases comme DBpedia [2] ou WikiData [3]. La Figure 5.7 illustre la modélisation de ces associations. Supposons que s_i soit une entité, instance de la classe S dans la figure, que l'on souhaite enrichir selon la connaissance locale en exemple. Le modèle d'enrichissement doit alors décrire les connaissances équivalentes dans les différentes sources externes (ici G1 et G2) selon le paradigme des triplets RDF.

5.3.2 Description des éléments du modèle

Nous décrivons maintenant les différents éléments qui composent un modèle d'enrichissement et nous commençons par la modélisation des connaissances locales.

Modélisation des connaissances locales

Pour rappel, les motifs de connaissances locaux (MBox) sont composés d'un ensemble de relations et propriétés (TBox) s'articulant autour d'une entité primaire (ABox). Par exemple, le motif d'*adaptation* est composé d'une relation qui lie une œuvre bibliographique (*ex.*, Da Vinci Code,

2. `http://wiki.dbpedia.org/about`
3. `https://www.wikidata.org`

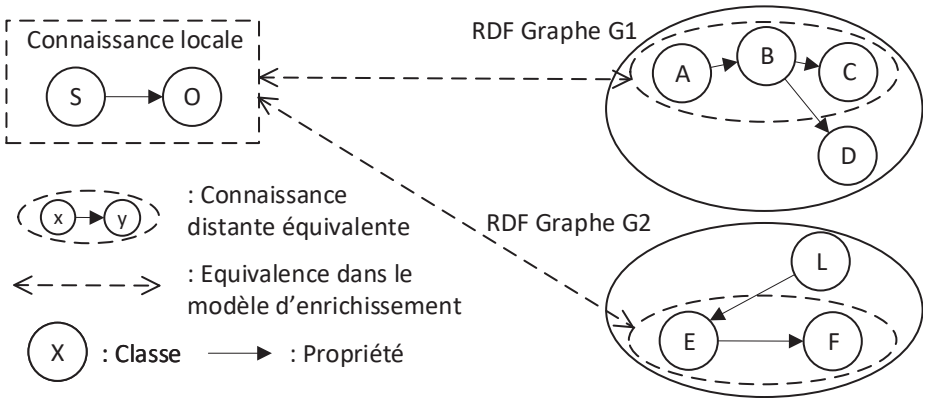


FIGURE 5.7 – Schématisation des relations d'équivalence entre la méta-représentation d'une connaissance locale et les entités RDF de sources distantes

le roman) à une autre œuvre (*ex.*, Da Vinci Code, le film). Ces relations et propriétés décrivent donc des informations remarquables sur l'entité à enrichir. Pour décrire ces informations dans le modèle d'enrichissement, nous réutilisons la notion de *Path* proposée par Nuzzolese, dans [96], et nous considérons que chaque motif de connaissances à enrichir est composé d'un ensemble de Paths.

Définition 5.3.2 (Path)**.** Dans un modèle (TBox) en graphe, un Path $P_{i,j} = [S_i, p, O_j]$ définit une relation p entre deux éléments du modèle S_i et O_j (*ex.*, classe, littéral), sous la forme d'un triplet RDF, telle que toutes les instances de ce modèle, suivant ce triplet, ont toujours le même type d'élément (*ex.*, `rdf:type`) pour S_i et pour O_j.

Exemple 5.3.1. Traductions d'une œuvre avec des *Paths*. Supposons que notre modèle local soit basé sur les classes de FRBR (*ex.*, Œuvre, Expression, Manifestation). La modélisation du motif de connaissances *Traduction*, c'est à dire les traductions d'une Œuvre, peut impliquer trois classes O_i pour l'Œuvre, E_{orig} pour l'Expression originale et E_{trad} pour l'Expression traduite et être représentée par trois *Paths*, un Path pour la réalisation originale, $Path_{realOriginale} = [O_i$,`rdaw:expressionOfWork`,$E_{orig}]$, un Path pour la deuxième réalisation, $Path_{realTraduction} = [O_i$,`rdaw:expressionOfWork`,$E_{trad}]$, et enfin un Path pour la relation de traduction $Path_{exprTraduction} = [E_{orig}$,`rdae:translatedAs`,$E_{trad}]$.

Modélisation des connaissances dans les sources externes

Nous admettons que les connaissances externes peuvent être structurées de manière différentes que les Paths locaux et que des processus supplémentaires d'interprétation des données peuvent être nécessaires. Pour représenter ces différences de modélisation, nous partons du principe que nous pouvons interroger les sources distantes selon le paradigme RDF, c'est à dire *sujet*, *predicat* et *object*. Nous utilisons alors la notion de *Property Path*, du W3C, à laquelle nous associons un ensemble de fonctions d'interprétations.

Définition 5.3.3 (Property Path)**.** Dans un modèle en graphe, un Property Path est un Path étendu $PP_{i,j} = [S_i, \delta\{p_i\}_k, O_j]$ retournant, au niveau instance, une séquence d'IRIs entre des entités de classe S_i et des entités de classe O_j. δ représente une expression associée à un ensemble de propriétés $\{p_i\}_k$ définissant la manière de *traverser* le graphe selon la sémantique de $\{p_i\}_k$. (*Dans un Property Path en SPARQL, δ est une expression régulière*). Dans la suite, nous utilisons les symboles [4] du W3C appliqués à une relation rel : $\hat{}rel$, rel_1/rel_2 et rel+ désignant respectivement une relation inverse de rel, une séquence de deux relations rel_1 puis rel_2 et un intervalle $[1, n]$ de rebonds sur une propriété rel.

4. `https://www.w3.org/TR/sparql11-property-paths/`

La grammaire associée aux Property Paths (voir [76]) contient les fonctions nécessaires pour naviguer dans un graphe RDF et spécifier la structure d'une connaissance à extraire. Par exemple, la fonction *inverse* définit si une relation à traverser entre deux classes du modèle est inverse par rapport au Path local. Nous illustrons les mécanismes élémentaires de ces fonctions avec l'exemple de deux modélisations différentes d'une même connaissance. La Figure 5.8 présente ces deux modélisations issues de deux modèles bibliographiques différents qui sont FRBR et FRBRoo. Cette connaissance, en langage naturel, peut s'interpréter par "l'Œuvre, ses créateurs et ses réalisations". Après avoir détaillé ces différences, nous proposons des exemples de *PropertyPath* pour chacun des deux modèles.

Sur la figure 5.8, nous observons d'abord la relation de *création* entre une Œuvre et son créateur. Dans FRBR, l'Œuvre est représentée par la classe O et, dans FRBRoo, par la classe $F15$. Le créateur est incarnée, dans FRBR, par la classe A (pour Agent) et, dans FRBRoo, par la classe P (pour Personne). La différence entre les modèles est que pour FRBRoo, il existe une classe intermédiaire $F27$ faisant le lien entre l'Œuvre $F15$ et une personne P. La notion de création implique donc deux propriétés dans FRBRoo contre une seule dans FRBR. Toujours dans la même figure 5.8, nous étudions la représentation des réalisations d'une Œuvre. Ici, FRBR utilise un lien direct entre l'Œuvre O et l'Expression E pour décrire cette connaissances. Dans FRBRoo, il existe une classe intermédiaire $F28$ faisant le lien entre l'Œuvre $F15$ et l'Expression $F2$. Ces exemples nous montrent qu'une même connaissance peut être décrite par un nombre différent d'entités entre deux modèles bibliographiques distincts. Dans cet exemple spécifique de FRBR et FRBRoo, FRBRoo utilise des entités supplémentaires car ce modèle décrit explicitement le processus de création et de réalisation d'une Œuvre là où ce processus est implicite dans FRBR. Il est donc crucial de tenir compte de ces différences dans les règles d'enrichissement.

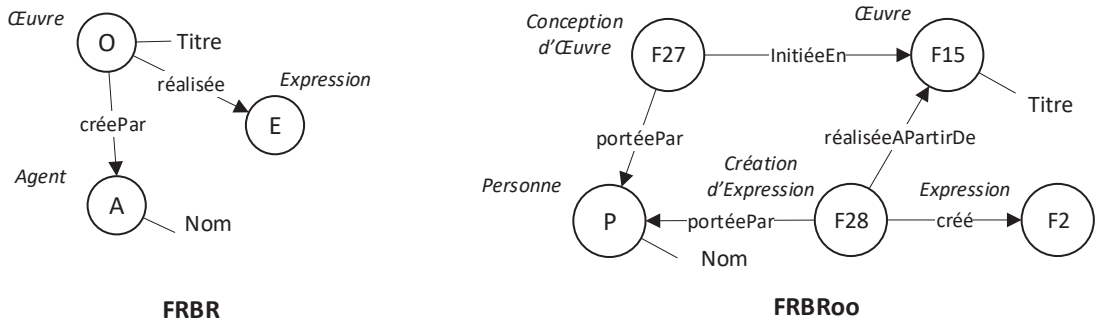


FIGURE 5.8 – Un même motif modélisé avec FRBR (à gauche) et FRBRoo (à droite)

Nous utilisons cet exemple de différences structurelles entre FRBR et FRBRoo pour définir plus formellement les éléments d'une règle d'enrichissement. Nous nous interessons à la connaissance suivante : l'Œuvre et ses réalisations. Pour rappel, cette connaissance nécessite la description d'un chemin entre trois entités F15 (Œuvre), F28 (Création d'Expression) et F2 (Expression) dans FRBRoo (ce qui représente un rebond supplémentaire par rapport à FRBR). De plus, la relation entre F15 et F28 est inversée, c'est à dire qu'elle n'existe que de F28 à F15 (`frbroo:réaliséeAPartirDe`) et doit donc être spécifiée dans la règle. Nous présentons maintenant deux exemples de Property Paths pour extraire cette information de FRBR et FRBRoo en tenant compte de ces différences.

Exemple 5.3.2. Réalisations d'une œuvre en Property Path Le Property Path pour FRBR peut s'écrire, [?O `rda:expressionOfWork` ?E], celui pour FRBRoo peut s'écrire [?O' `^frbroo:createdARealisationOf/frbroo:created` ?E'] où l'on considère que les entités ayant un `rdf:type` valant ?O sont équivalentes (`own:sameAs`) aux entités ayant un `rdf:type` valant

?O'. `^frbroo:createdARealisationOf` correspond à la relation inverse entre F28 et F15 et `/frbroo:created` au second rebond entre F28 et F2.

Association entre les modélisations locales et distantes

Le modèle d'enrichissement est une instance du méta-modèle des connaissances bibliographiques pour laquelle sont associées les modélisations locales (*Paths*) et distantes (*Property Paths*) des connaissances comme détaillées précédemment. Ces associations s'incarnent par des *correspondances d'enrichissement* qui lient les Paths de l'entité primaire avec les Property Path et fonctions d'interprétations associées aux différentes sources de données externes. Ces correspondances, rattachées aux différents motifs de connaissances, permettent la génération de requêtes (*ex.*, SPARQL) vers les sources distantes et l'intégration des données extraites dans la base locale. Nous rappelons que cette étape nécessite, comme pour la migration, un processus complémentaire de déduplication des entités qui sont intégrées incluant la fusion des propriétés.

Définition 5.3.4 (Correspondance d'enrichissement)**.** Une correspondance d'enrichissement est un quadruplet $\phi' =< Path_i, Src_j, PP_k, \{f_x\}_y >$ où $Path_i$ est un *Path* décrivant une relation de l'entité à enrichir, Src_j est une source de données externe, PP_k un *Property Path* représentant une équivalence de $Path_i$ dans la source Src_j et $\{f_x\}_y$ est un ensemble de fonctions d'interprétations appliquées à PP_k.

Les correspondances d'enrichissement viennent compléter les motifs de connaissances présentés dans ce chapitre afin de permettre un enrichissement des connaissances avec des sources de données externes et structurées. Nous présentons maintenant un cas d'usage, basé sur un exemple pratique, pour montrer l'utilisation des correspondances d'enrichissement de notre extension dans un contexte réel d'enrichissement d'une connaissance :

La connaissance que nous souhaitons enrichir est l'adaptation de romans en films avec la source de données DBPedia. Nous considérons que DBPedia ne représente pas le motif d'*adaptation* de manière explicite. Toutefois, cette information peut être déduite ou approchée pour être validée par un utilisateur. Dans cet exemple nous considérons le Path local $Path_{adaptFilm}$ = [O_{roman},`rdaw:adaptedAsWork`,O_{film}] selon lequel nous souhaitons intégrer les adaptations en films des romans issus de DBPedia, (si l'information est disponible). Quand une Œuvre et ses adaptations (*ex.*, en films) existent sur DBPedia, elles peuvent être regroupées de manière non-sémantique dans une ressource abstraite nommée *WikiPageDisambiguate*. Cette ressource RDF contient une liste basique de liens vers des ressources qui partagent théoriquement un label identique.
Par exemple, `dbpedia:The_Da_Vinci_Code_(disambiguation)` contient des liens vers les ressources décrivant le label "*The Da Vinci Code*" comme le roman, le film ou un jeu vidéo. Toutefois, cette liste de liens peut contenir d'autres ressources qui ne sont pas nécessairement des adaptations mais des critiques du roman ou complètement autre chose. Pour extraire les bonnes adaptations et reproduire le Path $Path_{adaptFilm}$ de notre base locale, nous devons analyser chaque ressource du *WikiPageDisambiguate* et retourner une information booléenne s'il s'agit bien d'une potentielle adaptation ou non. Pour cela nous proposons de créer une correspondance d'enrichissement dont le fonctionnement est illustré par la figure 5.9. Cette dernière schématise le processus d'interprétation de DBPedia pour combler l'absence de lien d'adaptation entre un roman (O'_A) et un film (O'_B).

Nous proposons de commenter la figure 5.9 en utilisant un exemple de *PropertyPath* $PP_{DBPediaDisamb}$. Ce dernier doit permettre de naviguer dans les ressources du *WikiPageDisambiguate* (W_d), depuis une entité de type *roman* (O'_A), avec deux fonctions d'interprétation respectivement $f_{typeAdapt}$ et $f_{commentAdapt}$. $f_{typeAdapt}$ retourne une valeur booléenne si la propriété `rdf:type` de la ressource

analysée contient un lien de type `yago:WikicatFilmsBasedOnBooks` et $f_{commentAdapt}$ retourne une valeur booléenne si la propriété `rdfs:comment` contient une information sur l'adaptation (ex : "adapted from [...] novel"). Nous créons ensuite la correspondance :

$\phi'_{DBPediaAdapt} =< Path_{adaptFilm}, DBPedia, PP_{DBPediaDisamb}, \{f_{typeAdapt}, f_{commentAdapt}\} >$
où $DBPedia$ est la source de données à interroger,

$PP_{DBPediaDisamb}$ = [`dbo:Book^dbo:wikiPageDisambiguates+/(rdfs:type|rdfs:comment)dbo:Film`], est le PropertyPath pour DBPedia et enfin $\{f_{typeAdapt}, f_{commentAdapt}\}$ correspond aux deux fonctions d'interprétations à appliquer sur DBPedia pour l'intégration de $Path_{adaptFilm}$.

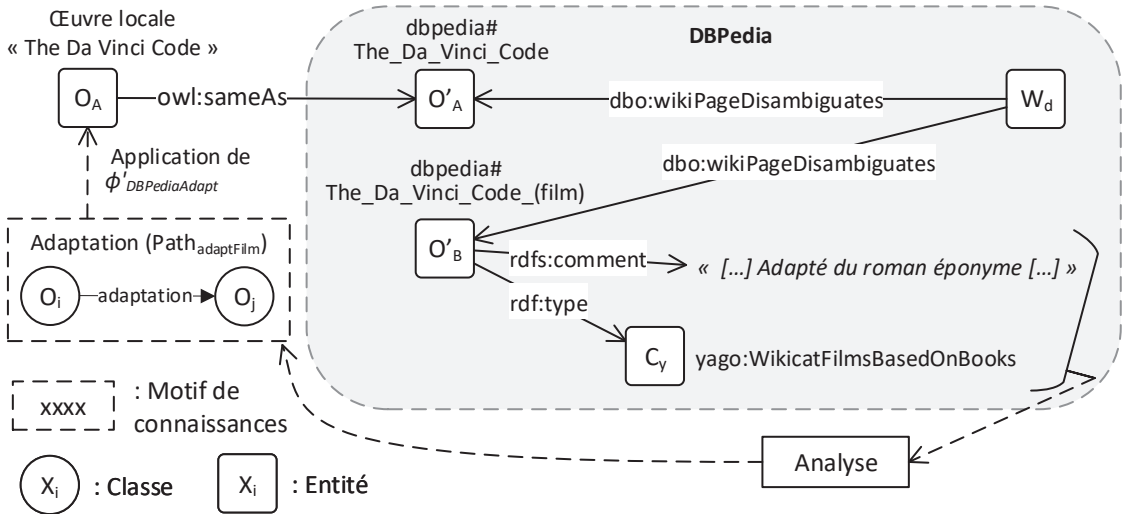


FIGURE 5.9 – Exemple d'enrichissement des adaptations d'une Oeuvre depuis DBPedia

Suivant les principes du web sémantique, il n'est pas conseillé de répliquer des données d'enrichissement sur chaque catalogue documentaire bénéficiant d'un tel processus mais plutôt de conserver uniquement les liens permettant l'interopérabilité entre les sources. C'est pourquoi, notre extension pour l'enrichissement peut être utilisée sans considérer l'intégration et la maintenance des données enrichies au sein des métadonnées locales. Par exemple, les données externes sont affichées à l'utilisateur au moment de la consultation des détails d'une Œuvre enrichie (en temps réel). Ce principe impliquant aux institutions d'interroger constamment certaines sources du web de données, différentes initiatives comme les Triple Pattern Fragments [138], la compression RDF [48] ou encore le projet LOD-a-lot [47] ont proposés des solutions intéressantes pour réduire les problèmes de performance lors de l'interrogation de sources de données distantes.

5.4 Méthodes de méta-modélisation

Nous avons défini les concepts essentiels pour la réalisation d'un modèle de migration et d'enrichissement bibliographique. Nous avons notamment détaillé la structure du modèle en arbre de motifs de connaissances. Dans cette section, nous étudions des cas pratiques de modélisation des connaissances en utilisant les principes de notre méta-modélisation. L'objectif est ici de fournir des clés pour appréhender la méta-modélisation des connaissances bibliographiques dans le contexte d'un processus de FRBRisation de notices.

5.4.1 Principes préliminaires de modélisation bibliographique

En préambule des cas de méta-modélisation des connaissances, nous présentons un ensemble de prérequis sur la construction des modèles de données bibliographiques.

Conception de nouveaux modèles bibliographiques

Différents modèles de données peuvent être conçus, à partir de FRBR, afin de répondre aux contraintes et spécificités des institutions qui les intègrent. Par exemple, une bibliothèque d'œuvres antiques n'utilisera pas le même modèle qu'un espace documentaire pour enfants. Les principaux modèles qui ont été expérimentés par la communauté documentaire ont été répertoriés et analysés dans [151] et [31]. Ces travaux ont permis de mettre en évidence les efforts nécessaires pour concevoir ces nouveaux modèles. Dans la suite de cette partie, nous illustrons cette tâche de modélisation avec un exemple représentatif.

Considérons une institution documentaire qui souhaite créer une base de connaissances permettant la gestion de dossiers qui contiennent des journaux et des articles scientifiques. L'institution décide de s'inspirer des principes FRBR pour élaborer le modèle de cette future base de connaissances. La première étape de ce travail consiste à adapter le modèle conceptuel FRBR aux contraintes de l'institution. Nous prenons trois exemples pour illustrer ce travail d'adaptation : (1) *Comment distinguer un dossier d'une Œuvre FRBR ?* (2) *Comment représenter les affiliations des auteurs scientifiques avec la classe Agent ?* (3) *Comment gérer les journaux ayant plusieurs éditions avec la classe Manifestation ?* Sur la Figure 5.10, nous présentons un exemple de solution pour ces questionnements. Ici, trois classes *Dossier*, *Scientifique* et *Édition* ont été dérivées des classes conceptuelles de FRBR. La Table 5.1 donne une description de ces trois nouvelles classes.

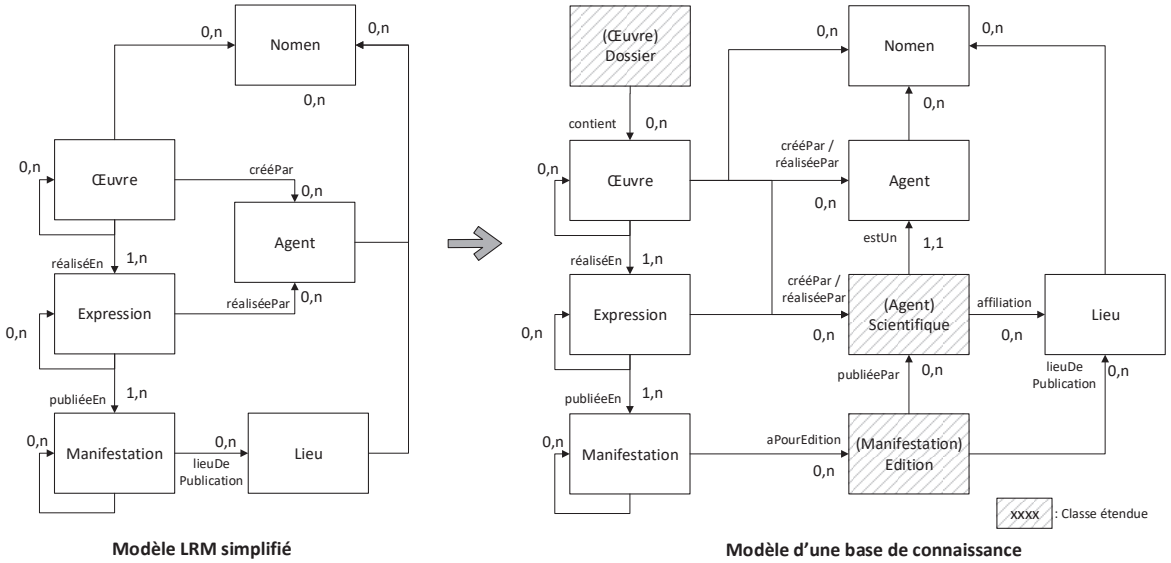


FIGURE 5.10 – Évolution d'un modèle (TBox) pour intégrer les contraintes spécifiques d'un domaine

Dans la Figure 5.10, un Dossier est une spécialisation de la classe Œuvre permettant ainsi de faciliter la distinction entre les Œuvres intellectuelles (*ex.*, articles, journaux) et les dossiers proposés par l'institution. Toutefois, chaque Dossier hérite directement de la classe abstraite Œuvre. Toujours dans cet exemple, les *Agents*, qui écrivent les articles scientifiques, peuvent avoir différentes affiliations. Ces dernières étant représentées par la classe *Lieu* dans notre exemple,

Entité	Définition	Remarque
Dossier	Spécialisation d'une œuvre comme une œuvre d'agrégation	*Permet de contenir des agrégations d'œuvres tout en gardant distincte la notion d'œuvre intellectuelle.*
Scientifique	Contextualisation d'une classe Agent ayant un contexte variable.	*Permet de représenter des responsabilités dans des contextes spécifiques comme un ordre d'apparition ou une affiliation.*
Édition	Contextualisation d'une manifestation pour distinguer plusieurs éditions	*Utilisé dans le cas de multiples publications (manifestations) ayant des lieux et/ou dates différentes.*

TABLE 5.1 – Classes étendues

l'*Agent* est contextualisé en un *Scientifique* pour conserver ces informations d'affiliations. C'est le même principe pour l'*Édition* qui est une contextualisation de la *Manifestation* pour distinguer des *Lieux* de publication différents en cas de rééditions (*ex.*, pour les journaux). Cet exemple de la Figure 5.10 s'inspire plus généralement des méthodologies de modélisation des connaissances comme dans [102] ou [7], où sont présentés des exemples plus complets de modélisation dans l'univers documentaire. La tâche essentielle dans ces processus, que nous illustrons dans notre exemple, est appelée *conceptualisation*.

Définition 5.4.1 (Conceptualisation)**.** Lors de la création d'un modèle métier, la conceptualisation consiste à spécialiser et/ou contextualiser les concepts abstraits pour permettre la représentation des informations du métier. Le résultat de cette tâche s'incarne en de nouvelles classes et propriétés qui précisent la sémantique des concepts initiaux et qui permettent de répondre aux contraintes et spécificités du métier.

Comme pour les auteurs de [72], qui ont récemment proposé des outils pour assister ce travail de conceptualisation, un des enjeux de cette thèse est de faciliter la création de nouveaux modèles de données bibliographiques en respectant les enjeux de qualité et les contraintes métier des institutions documentaires. Dans cette partie, nous avons présenté des principes élémentaires de la modélisation de métadonnées bibliographiques avec FRBR. Dans la dernière partie de ce chapitre, nous abordons des perspectives plus expérimentales sur la création de bases de connaissances bibliographiques, selon ces principes.

Représentation des relations bibliographiques avancées

Le travail de conceptualisation, d'éléments du modèle FRBR, permet de répondre aux contraintes directes d'une institution concernée par l'adoption de ces nouveaux standards. Cependant, la valorisation du patrimoine bibliographique implique d'aller plus loin dans l'effort de modélisation, en intégrant notamment des relations bibliographiques plus avancées (*cf.* section précédente). En effet, des travaux récents ont montré les bénéfices, sur l'expérience utilisateur, que peut apporter la modélisation des familles bibliographiques, de manière explicite, dans de nouvelles bases de connaissances FRBR [90, 3]. Par exemple, la représentation explicite des adaptions audio et des traductions d'une Œuvre offre de nouvelles possibilités pour faciliter la recherche de contenus bibliographiques par des personnes malentendantes. C'est pourquoi, la communauté documentaire, avec des travaux comme [149] ou [22], étudie la possibilité de modéliser ces familles bibliographiques avec les classes et propriétés de FRBR. Toutefois, pendant la rédaction de cette thèse,

ce type de modélisation n'en est qu'à ses prémices. En effet, il n'existe encore que très peu de bases de données, y compris dans le web de données, qui utilisent le potentiel de FRBR et du Web Sémantique pour modéliser des relations et des familles bibliographiques avancées.

Dans la suite de cette section, nous présentons et discutons quelques pistes pour la modélisation de relations bibliographiques dans une base de connaissances intégrant les principes FRBR. La Figure 5.11 présente quatre exemples que nous qualifions de *motifs de connaissances* et qui s'inspirent des familles bibliographiques connues [133]. Ici, les classes O_i sont des Œuvres, les classes E_i, des Expressions, les classes M_i, des Manifestations et enfin les classes A_i, des Agents.

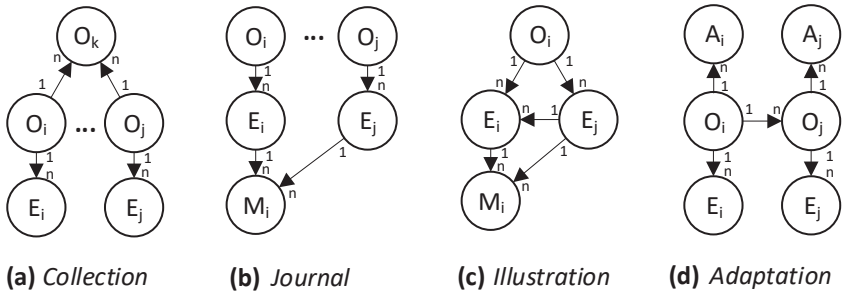


FIGURE 5.11 – Exemples de motifs courants de relations bibliographiques

— **Motif (a)**. Sur la Figure 5.11, une collection (ou agrégation) O_k de plusieurs œuvres O_i peut être modélisées par le motif (a). Ici, chacune des Œuvres agrégées O_i possède sa propre réalisation (Expression) E_i.

— **Motif (b)**. Une œuvre périodique (*ex.*, un journal) O_i, dans lequel sont publiées différentes œuvres (*ex.*, des articles) O_j peut être modélisé avec le motif (b). Dans ce motif, ce sont les réalisations E_j des œuvres O_j qui sont publiées dans la Manifestation M_i du journal.

— **Motif (c)**. Une œuvre O_i, qui a été augmentée peut être modélisée par le motif c où les éléments ajoutés (*ex.*, illustrations) E_j sont liés à la réalisation principale E_i de l'Œuvre augmentée O_i. Si ces illustrations n'ont pas leur propre publication, alors elles sont intégrées dans la même Manifestation M_i que la réalisation principale E_i.

— **Motif (d)**. Une œuvre (*ex.*, un livre) O_i, adaptée en une autre œuvre (*ex.*, un film) O_j peut être modélisée avec le motif d.

La standardisation des relations entre les classes de modèles conceptuels comme FRBR, qui forment les familles bibliographiques, est d'autant plus complexe qu'une même information peut être modélisée de différentes manières selon les pratiques des institutions. Aalberg *et al*, dans [5], ont mis en avant certains enjeux sur la modélisation des relations bibliographiques. Par exemple, une augmentation d'une Œuvre (*ex.*, illustration), doit elle-être modélisée comme une nouvelle Œuvre à part entière ou comme l'Expression de l'Œuvre augmentée ? Tillet [132] a proposé des pistes de réflexion sur la méthodologie de modélisation des familles bibliographiques. La Figure 5.12 présente une illustration de Tillet montrant si une dérivation d'Œuvre (*cf.*, Section précédente) doit être modélisée avec une nouvelle Œuvre liée à Œuvre dérivée (*ex.*, adaptation en film), ou avec une nouvelle Expression de l'Œuvre dérivée (*ex.*, traduction).

Dans ce préambule, nous avons décrit certains principes préliminaires pour la modélisation des connaissances bibliographiques. Dans la suite de cette section, nous détaillons différents cas de méta-modélisation caractéristiques des projets de construction de bases de connaissances documentaires.

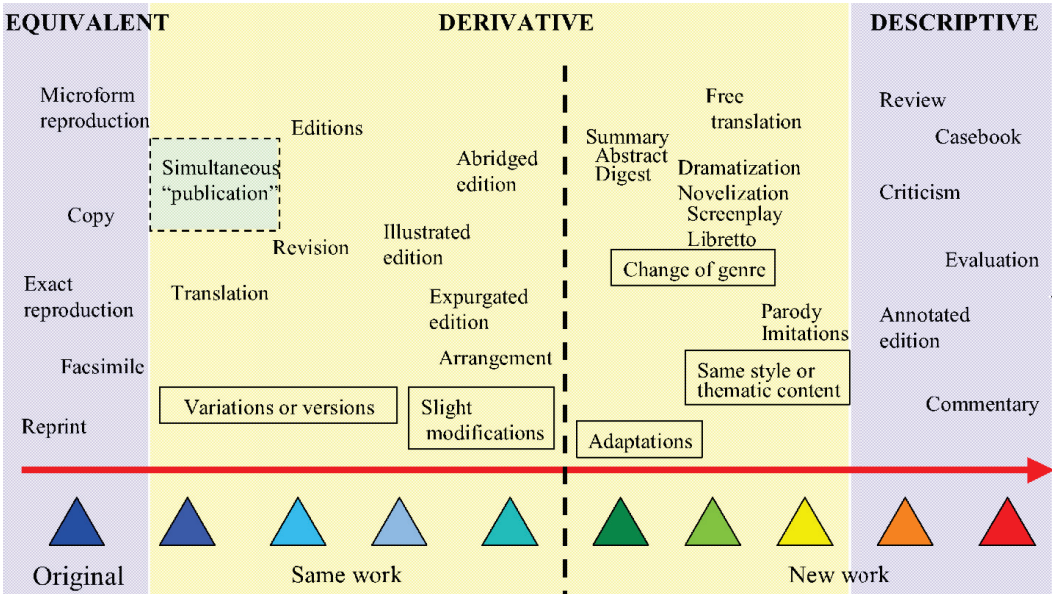


FIGURE 5.12 – Famille des dérivations d'œuvres

5.4.2 Modélisation élémentaire

Un catalogue bibliographique contient en majorité des notices de ressources isolées ou ne décrivant pas de contexte intellectuel ou éditorial avancé. Ces notices, les plus "simples", peuvent être transformées avec uniquement les correspondances et motifs les plus élémentaires du méta-modèle des connaissances. Ces derniers sont en principe communs avec l'ensemble du catalogue.

Définition 5.4.2 (Motif de connaissances élémentaire)**.** Un motif de connaissances élémentaire est un graphe orienté de classes, issues du modèle cible, où chaque nœud a au plus un prédécesseur. (Nous omettons volontairement les notions de correspondances de migration ou d'enrichissement ici car c'est la modélisation qui nous intéresse).

Exemple 5.4.1. Le premier groupe du modèle conceptuel FRBR est un motif élémentaire (*cf.*, chapitre 2). La représentation de ce groupe comme un motif de connaissances consiste en une Œuvre O_i (donnant lieu à l'entité primaire) qui est créée par un ou plusieurs Agents $\{A_i\}_j$ et qui est réalisée en une ou plusieurs expressions $\{E_i\}_j$, elles-mêmes publiées en une ou plusieurs Manifestations $\{M_i\}_j$. La Figure 5.13 présente l'application de ce motif sur une notice simple décrivant le livre *'Little Brother'* écrit par R. Enthoven.

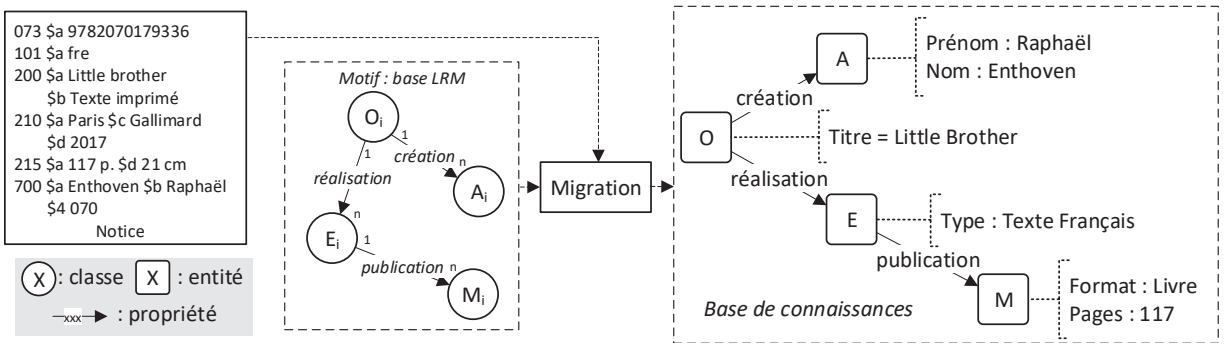


FIGURE 5.13 – Motif FRBR élémentaire appliqué à la notice A

La migration de notices dites "simples" implique de décrire des correspondances pour le motif élémentaire. Par exemple, nous proposons une définition de la correspondance, issue de la

Figure 5.13, qui est relative à la réalisation de la classe Œuvre O_i en une classe Expression E_i telle que $\phi_e =< \{101\$a, 200\$b\}, \{(réalisation(O_i) \rightarrow E_i)\}, < typeExpr, codeLangue >\}$ où $101\$a$ et $200\$b$ sont les champs concernés, $\{(réalisation(O_i) \rightarrow E_i)\}$ représente la part du motif à construire (composé des classes *Oeuvre* et *Expression* et de la relation *réalisation*) et $typeExpr(format, langue)$ et $codeLangue(code)$ sont des fonctions permettant d'interpréter les champs concernés, avec ici $typeExpr(200\$b, codeLangue(101\$a)) = Texte\ Français$.

Cet exemple illustre la manière de modéliser des motifs élémentaires de connaissances avec notre méthodologie de méta-modélisation.

5.4.3 Modélisation contextualisée

Les modèles de notices bibliographiques permettent la description d'entités d'une même classe dans différents contextes. Par exemple, deux entités de classe Agent peuvent être liées à une même entité de classe Œuvre, mais en ayant un rôle différent (comme *auteur* ou *traducteur*). La construction des motifs, dans le modèle de migration, doit donc permettre de distinguer ces classes identiques mais décrivant une entité différente. C'est pourquoi, notre instance du méta-modèle permet une contextualisation des classes dans les motifs de connaissances. Cette contextualisation s'incarne par une appellation (label) qui peut varier entre des classes identiques d'un même motif.

Par exemple, sur la figure 5.14, le motif (a) décrit une Œuvre Oi, réalisée en une ou plusieurs Expressions E_i (*ex.*, texte en français, audio en anglais). Nous illustrons ces multiples Expressions par une cardinalité 1-n. Si ces différentes Expressions sont modélisées au sein d'un même motif du méta-modèle, alors il convient de les distinguer comme ici où E_i est une Expression distincte de E_j. Selon le même principe, le motif (b) décrit une Œuvre O_i pouvant être créée par un ou plusieurs agents $\{A_i\ ..\ A_j\}$. Le motif (c) décrit une réalisation E_i pouvant être publiée en une ou plusieurs Manifestations $\{M_i\ ..\ M_j\}$ différentes (*ex.*, livre de poche, version numérique).

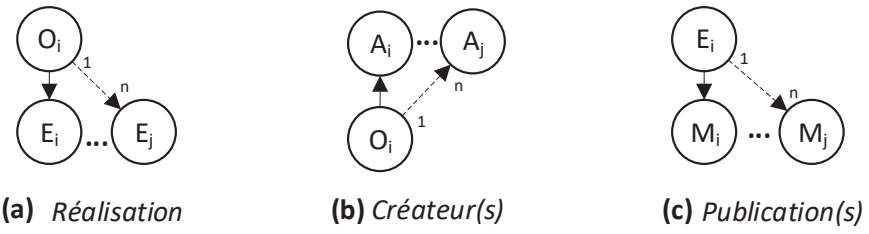


FIGURE 5.14 – Exemples de motifs contextualisés

Contextualisation : éviter la redondance des propriétés

Nous avons vu que les entités produites dans la base de connaissances sont créées à partir des classes et propriétés du modèle cible selon certaines conditions. Ces dernières dépendent des spécificités des notices en entrée. C'est pourquoi, les classes qui sont modélisées sont décrites avec leurs propriétés qui peuvent être des attributs ou des relations (*ex.*, une *Œuvre* avec son attribut *titre*) et sont soumises à des conditions qui peuvent s'appliquer, dans l'instance du méta-modèle, à l'échelle d'une classe ou d'un motif (contenant plusieurs classes). Par exemple, une condition qui évalue la présence d'un champ de *langue traduit*e et un code de fonction *traducteur*, dans une notice, peut servir à activer le motif *traduction* (*cf.*, figure 5.5). Cependant, notre mécanisme de contextualisation permet la représentation de plusieurs classes identiques faisant référence à des entités différentes (*ex.*, la *traduction* implique au moins deux classes Expressions). Le problème ici concerne la redondance dans les correspondances de migration. En effet, il s'agit d'éviter la

répétition des attributs d'une classe (*ex.*, Expression), si cette dernière est déjà représentée plusieurs fois. De même, nous ne souhaitons pas répéter les différentes conditions qui permettent de déclencher l'instanciation d'une classe si cette dernière est réutilisée dans plusieurs motifs.

C'est pourquoi, nous utilisons un mécanisme supplémentaire, dans l'instance du méta-modèle, afin d'éviter toute redondance dans les correspondances de migration. Ce mécanisme repose sur l'héritage des attributs et conditions entre les motifs grâce à la structure en arbre de l'instance du méta-modèle. Pour rappel, les nœuds de l'arbre constituent différents motifs de connaissances. Ce mécanisme d'héritage permet de simplifier la lisibilité des motifs et de réduire les calculs inutiles selon la structure globale de l'arbre. Nous illustrons maintenant ces bénéfices au travers de deux exemples représentant des situations courantes dans les catalogues bibliographiques.

Exemple 5.4.2. Différentes relations vers une même classe. La Figure 5.15 illustre la création d'une instance du méta-modèle pour la migration d'une notice ainsi qu'un exemple de base de connaissances produite. La notice décrit une bande dessinée, écrite par Lupano W. et illustrée par Cauuet P. Nous avons donc deux Agents (zone 700) ayant une fonction différente avec l'œuvre, caractérisée par les codes ($4), respectivement *690* et *440* de chaque zone 700. Nous souhaitons, à l'issue de la migration, intégrer le *nom* et le *prénom* de chaque Agent dans la base de connaissances. L'instance proposée utilise trois motifs du méta-modèle qui sont respectivement *Réalisation*, *Agent scénariste* et *Agent illustrateur*. Elle intègre aussi un motif générique qui est ajouté pour éviter de répéter la description des attributs *nom* et *prénom* dans les deux motifs concernés. Ici, le mécanisme d'héritage de ces attributs permet d'assigner correctement les noms et prénoms des deux Agents dans la base de connaissances. Cette notion de motif générique est simplement une manière d'exploiter l'héritage entre les motifs dans l'arbre du modèle de migration.

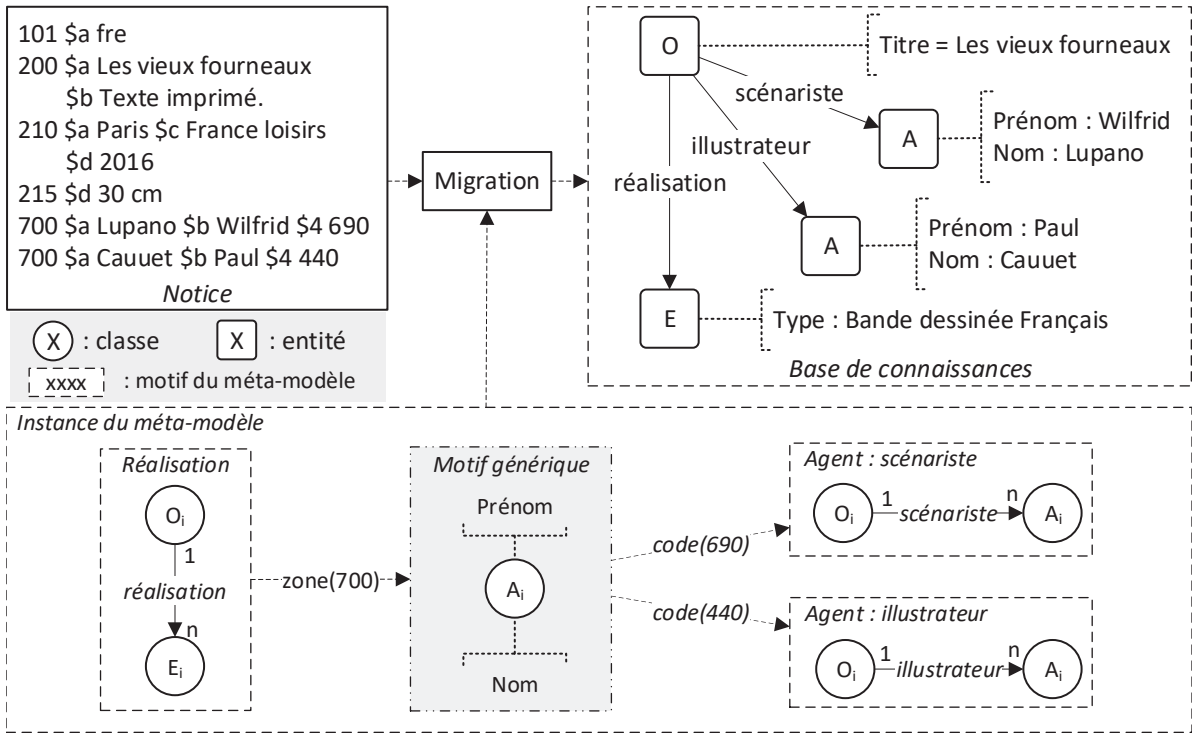


FIGURE 5.15 – Modèle de migration avec héritage des attributs de classe

Exemple 5.4.3. Différents attributs pour une même classe Nous présentons maintenant l'exemple d'une relation bibliographique courante, la traduction d'une œuvre. La modélisation de

cette relation implique la création de deux entités de la classe *Expression* (réalisation de l'œuvre) pour représenter les deux contextes *Expression Traduite* et *Expression Originale*, relatives aux deux langues différentes. La Figure 5.16 présente une notice relative à la traduction en français (*fre*) du roman *The Da Vinci Code*, initialement en anglais (*eng*). Dans la base de connaissances produite, nous souhaitons distinguer le contexte et la langue de chaque Expression de l'Œuvre. L'instance proposée intègre donc trois motifs du méta-modèle qui sont respectivement *Réalisation*, *Traduction* et *Traducteur*. Ici, l'instance intègre également une condition sur le motif *Traduction*, symbolisée par la fonction booléenne *trad*(). Cette fonction analyse une notice pour détecter la présence d'une traduction. Dans l'arbre des motifs de l'instance, le motif *Traducteur* est alors placé comme fils du motif *Traduction* afin d'hériter de cette condition.

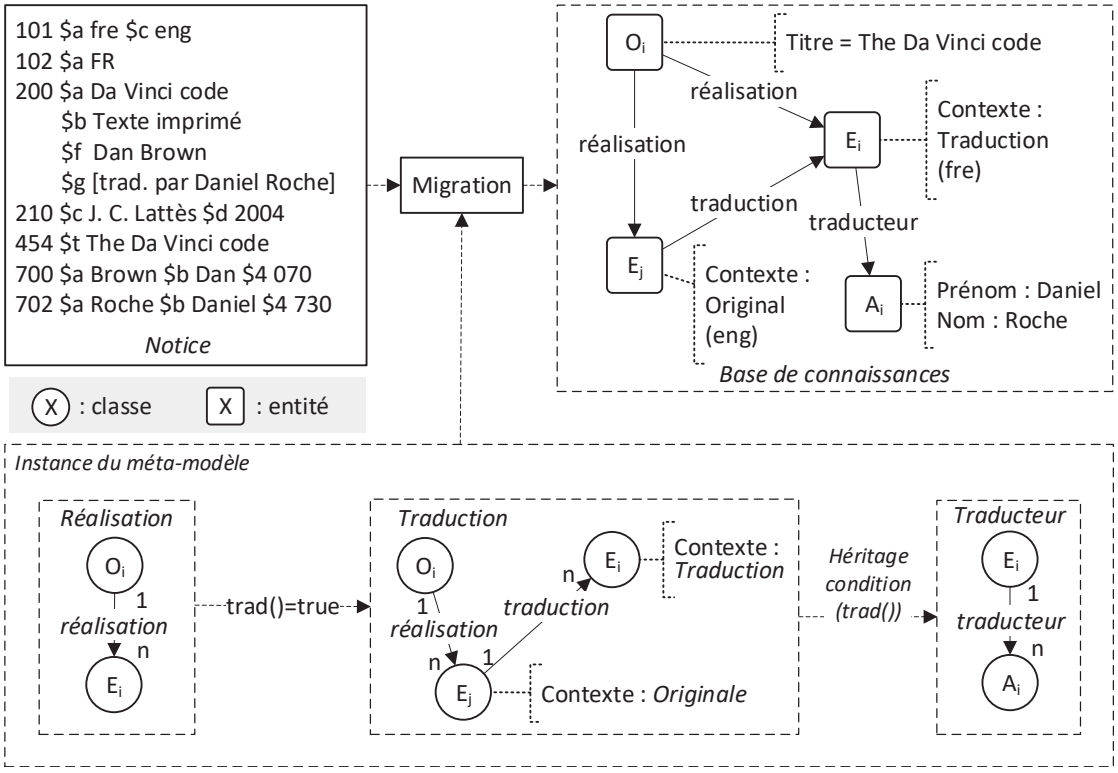


FIGURE 5.16 – Modèle de migration avec héritage d'une condition de motif

Nous avons montré des mécanismes essentiels dans notre méthodologie de modélisation des connaissances. Ces derniers, comme la contextualisation ou encore l'héritage, permettent de considérer certaines relations bibliographiques avancées comme la *traduction* d'une Œuvre. Cependant, d'autres relations issues des agrégations ou des dérivations d'œuvres (*cf.*, chapitre 2) nécessitent des mécanismes supplémentaires pour être correctement modélisées. Il s'agit notamment des relations qui impliquent des informations issues de plusieurs notices du catalogue.

5.4.4 Modélisation multi-notices

Nous étudions le cas de modélisation où les motifs de connaissances impliquent des correspondances entre plusieurs notices. Pour illustrer ce type de modélisation, la Figure 5.17 contient deux exemples de notices bibliographiques relatives à l'œuvre de Dan Brown, *The Da Vinci Code*. La *notice A* décrit la monographie originale en anglais. La *notice B* décrit le film DVD en français, adapté de l'œuvre originale. Dans ces exemples nous pouvons observer la présence de relations bibliographiques comme une *augmentation* (expression illustrée) dans la *notice A* et une *adaptation* (en film) dans la *notice B*. La Figure 5.17 détaille le processus d'application du

modèle de migration sur ces deux notices.

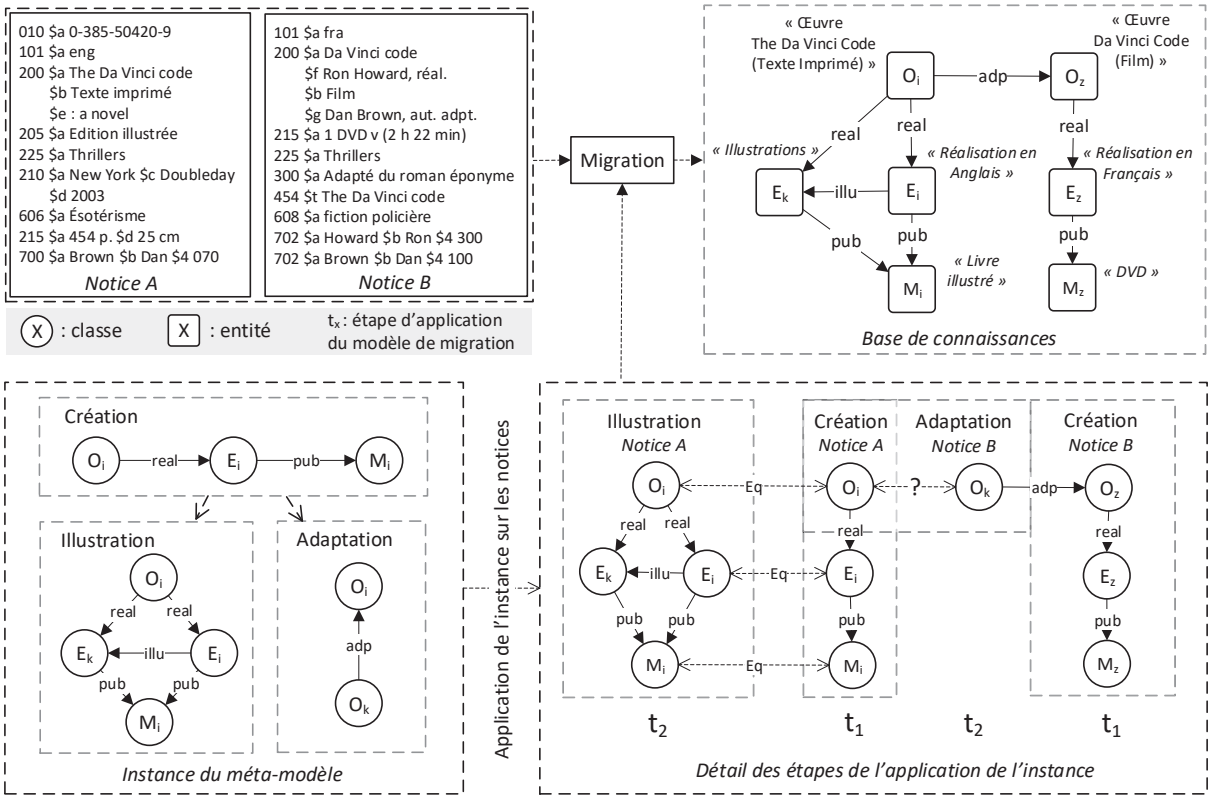


FIGURE 5.17 – Interprétation de deux notices liées

En entrée du processus de modélisation nous avons deux notices A et B. Afin de modéliser les connaissances contenues dans ces notices, l'instance proposée utilise trois motifs du méta-modèle qui sont les motifs *Création*, *Illustration* et *Adaptation*. Le premier, *Création*, permet de construire la branche initiale FRBR décrivant qu'une œuvre bibliographique est réalisée en une Expression, elle-même publiée en une Manifestation. Le deuxième motif, *Illustration*, décrit la dérivation de l'Œuvre (O_i) en deux Expressions (E_k et E_i) d'une même Manifestation (M_i) avec E_k représentant la réalisation de l'illustration. Le troisième motif, *Adaptation*, décrit une relation entre deux œuvres (O_i et O_k) où O_k, qui représente l'adaptation d'une œuvre, est dérivée de O_i.

Dans cet exemple, figure 5.17, nous proposons le détail du parcours de l'arbre des motifs pour la migration des deux notices. Chaque étape du parcours est numérotée par un temps t_x. Cette illustration permet donc d'observer la création des différents éléments de chaque motif pour les notices concernées. Avec le mécanisme de contextualisation des classes, la figure 5.17 détaille aussi les relations d'équivalences qui correspondent aux classes qui doivent être fusionnées pour former les entités de la base de connaissances. Par exemple, l'entité Œuvre, représentée par la classe O_i au temps t_1 est équivalente à celle représentée par la classe O_i au temps t_2. Dans la base de connaissances, une seule entité contient les attributs relatifs à O_i pour t_1 et t_2. La subtilité de cet exemple est que, lors de l'application des motifs, nous ne sommes pas en mesure d'évaluer si O_i, O_k mènent à la même œuvre dans la base de connaissances, car venant de notices différentes. Nous symbolisons cette incertitude par $<-?->$ sur la Figure 5.17. Cette information doit être déduite par une phase supplémentaire qui évalue l'équivalence des entités provenant de notices différentes (phase décrite au chapitre suivant). Dans notre exemple, O_z, le film, est adapté de O_k qui est en réalité le livre original *The Da Vinci Code* (information présente dans les champs 300$a et 500$a de la notice B), représenté par O_i. Ainsi, dans la base de connaissances, O_i et O_z

sont fusionnés pour ne former qu'une seule entité, le livre original.

Les exemples présentés précédemment permettent d'explorer la modélisation de motifs courants dans les catalogues bibliographiques comme l'*illustration*, la *traduction* ou l'*adaptation*. Ces motifs sont illustrés dans les notices relatives à l'œuvre *The Da Vinci Code*. En guise de synthèse de ces exemples, nous présentons, sur la figure 5.18, une proposition de base de connaissances issue de la migration de ces notices avec les motifs détaillés jusqu'ici. Dans la suite de cette partie nous présentons des exemples plus spécifiques et moins courants dans la modélisation des connaissances bibliographiques.

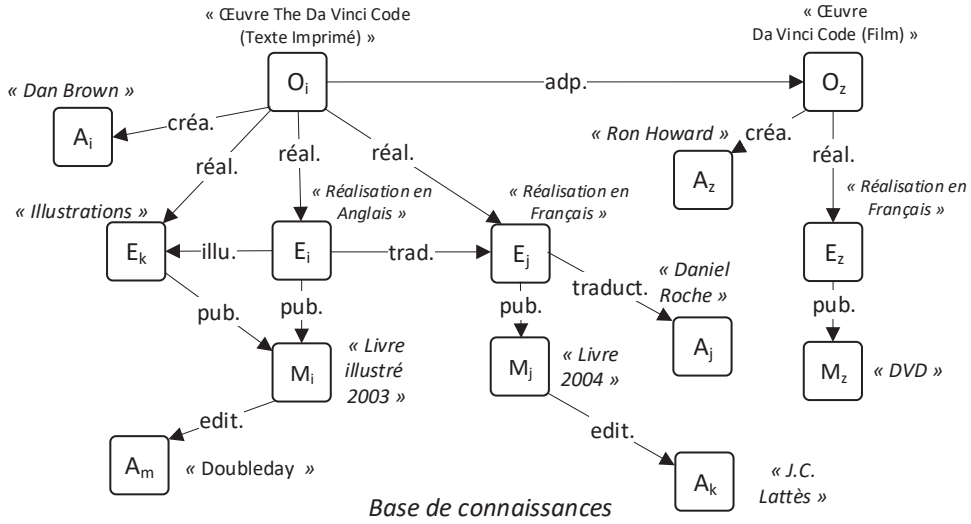


FIGURE 5.18 – Migration des motifs *Illustration*, *Traduction* et *Adaptation*

5.4.5 Modélisation multi-niveaux

Dans l'interprétation d'un catalogue bibliographique, certains motifs de connaissances à extraire peuvent être spécifiques au contexte du catalogue à migrer et nécessiter une modélisation particulière. Un exemple de modélisation spécifique concerne les motifs dont la sémantique doit dépendre d'au moins deux niveaux de relations entre les classes. C'est à dire que chaque motif inclut au minimum 3 classes distinctes. Nous illustrons ce type de motifs avec deux exemples relevés d'un catalogue bibliographique scientifique dans le domaine médical. La figure 5.19 illustre un premier exemple de motif à 2 niveaux avec une notice décrivant un journal scientifique.

Sur la figure 5.19, nous nous interessons aux éditions d'un journal scientifique. Une édition est en principe décrite par une publication, comme une *Manifestation* FRBR, réalisée par un ou plusieurs *Agents* FRBR qui sont des éditeurs. Ici, le journal possède plusieurs éditions par des éditeurs différents (*Slack* et *Cambridge*), à des périodes différentes. L'information de période d'édition étant initialement placée sur la classe *Manifestation* M_i, il convient ici de distinguer les deux périodes du journal par des classes *Editions* Ed_i, enfant de la *Manifestation*. Chaque éditeur peut ainsi être accroché à la bonne Édition Ed_i. Nous proposons donc le motif *Editions multiples*, pour intégrer cette notion d'*Édition*.

La migration de cette notice, selon ce motif nécessite un traitement spécifique pendant la lecture des champs de la notice. En effet, si une *Édition* peut être déduite d'une zone (ici 210), un *Agent*,

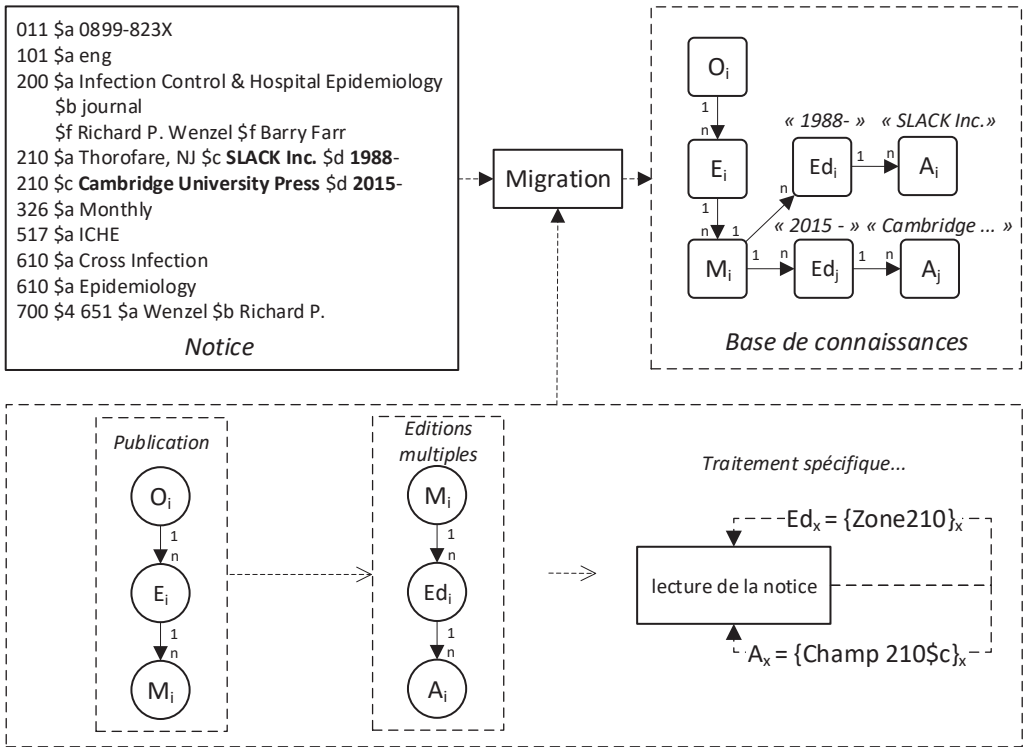


FIGURE 5.19 – Modélisation et migration pour des éditions multiples

lui, est représenté par un champ de la zone (ici 210$c). Le processus de traitement des notices, lors de la migration, doit donc tenir compte de ces champs, menant à des entités et non pas à des propriétés, afin de migrer correctement les informations.

La Figure 5.20 illustre un deuxième exemple de motif spécial. Dans cet exemple, nous migrons deux notices qui décrivent des articles scientifiques ayant un auteur en commun, *Rhee Y.*. La particularité de cet exemple est que, pour la notice A, l'auteur *Rhee Y.* a une affiliation scientifique correspondant à l'organisme *University Medical Center*. Dans ce contexte, l'enjeu pour l'instance du méta-modèle est de considérer, pour une œuvre donnée, les auteurs ayant une affiliation (que nous appelons *auteurs scientifiques*), et ceux qui n'en n'ont pas. Dans la figure 5.20, nous proposons un exemple d'instance utilisant les motifs *Création*, *Créateur* et *Créateur scientifique*. Ce dernier motif, activé si une affiliation est décrite dans la notice, introduit une classe SA_i qui permet de lier un auteur et une affiliation pour une œuvre donnée. La base de connaissances proposée dans l'exemple montre l'application de ce motif sur les deux notices.

Dans cette section plus pratique, nous avons présenté différents mécanismes qui viennent compléter notre modèle de migration et d'enrichissement dans le cadre de notre méthodologie de modélisation pour la transformation de catalogues documentaires. Certains mécanismes appliqués à notre modèle (*ex.*, contextualisation des classes, héritage des attributs) permettent d'encapsuler les correspondances de migration afin de tenir compte des relations bibliographiques avancées d'un catalogue. Les autres mécanismes présentés plus haut permettent d'intégrer des cas plus spécifiques lors de la migration (*ex.*, lecture des cardinalités des zones et champs) de notices particulières afin d'interpréter correctement les informations du catalogue et de pouvoir modéliser clairement les connaissances.

Nous rappelons que cette méthodologie a été crée comme un complément des outils de migration et d'enrichissement de la communauté documentaire. Elle peut être adaptés aux besoins spéci-

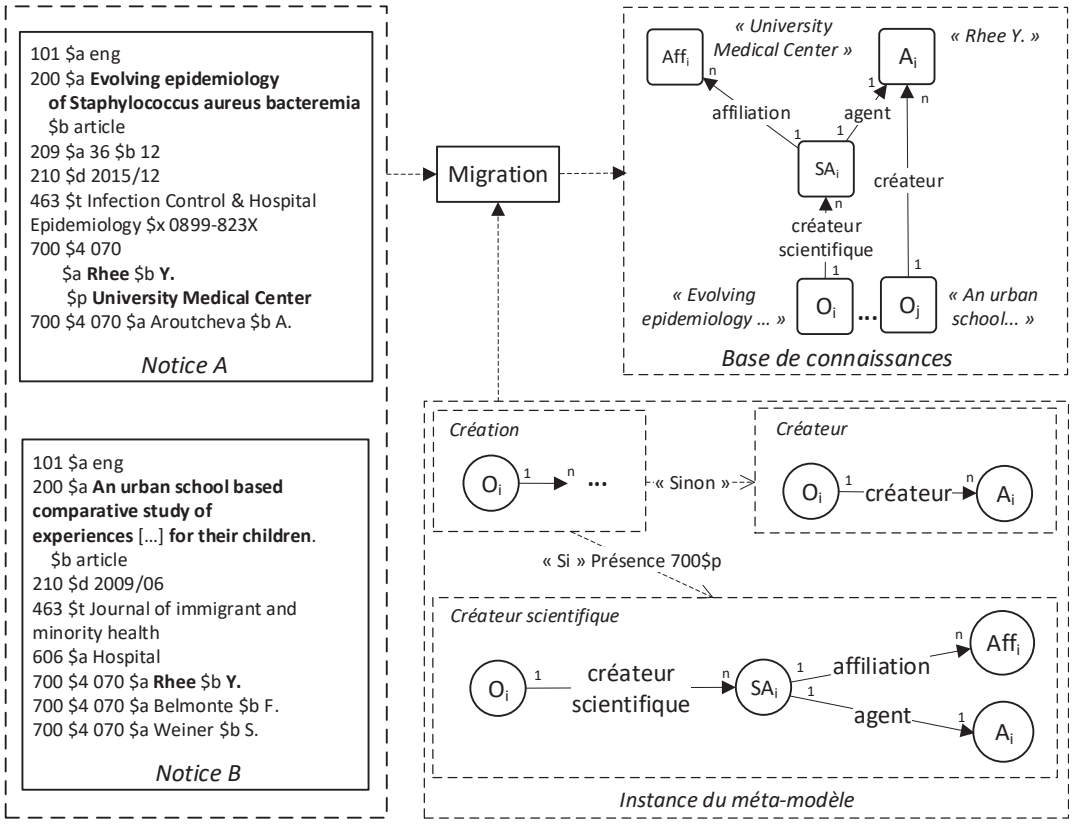


FIGURE 5.20 – Modélisation avec une entité contextualisée

fiques d'un projet et a pour principale ambition de faciliter la réalisation de projets d'adoption des principes du Web Sémantique dans les institutions culturelles.

5.5 Réutilisation des correspondances et motifs

Un des objectifs de notre méta-modélisation est de produire des règles qui soient réutilisables dans différents projets afin de réduire les coûts de paramétrage. Dans le cadre de la FRBRisation par exemple, la réutilisation des règles pour plusieurs catalogues doit permettre un gain de temps dans la préparation d'un nouveau projet de migration. En effet, la contextualisation de classes et l'héritage des conditions permet notamment d'apporter des modifications à un modèle de migration en identifiant rapidement le motif concerné et en évitant d'introduire de la redondance dans les règles. Toutefois, ces avantages peuvent dépendre de la modélisation des motifs d'un projet et des spécificités du catalogue. C'est pourquoi, dans cette section, nous évaluons le potentiel de réutilisation des motifs de connaissances issus de notre méthodologie. L'objectif est d'observer si les motifs de connaissances qui sont présentés dans les chapitres 4 et dans les sections précédentes de ce présent chapitre 5 peuvent être réutilisés dans différents contextes bibliographiques.

Pour réaliser cette étude, nous évaluons un même ensemble de règles de migration, créées selon notre méthodologie, sur différents jeux de données. Nous avons rassemblé 3 jeux de données du monde réel couvrant des types de ressources documentaires différents comme des ressources pédagogiques, scientifiques ou de lecture pour le grand public. Le premier jeu de données, appelé LUOM, correspond à 200.000 notices bibliographiques extraites aléatoirement d'une bibliothèque universitaire Américaine. Ce jeu de données comporte des notices décrivant des ressources variées, du monde scientifique à la lecture publique, avec une dominance pour les ressources pédagogiques. Le second jeu de données, HCL, consiste en 100.000 notices extraites d'une bibliothèque médi-

cale. Ces notices sont donc exclusivement spécialisées dans le domaine scientifique. Enfin, le jeu de données que nous appelons BNF contient 200.000 notices extraites aléatoirement de la Bibliothèque Nationale de France. Les notices BNF peuvent décrire des ressources de types variés avec une dominance pour la lecture publique. Le jeu de règles de migration que nous avons créé est basé sur notre approche de méta-modélisation présentée dans ce chapitre et peut être téléchargé en ligne au format RDF [5]. Il se compose d'un arbre de 63 motifs qui encapsulent un total de 368 correspondances de migration. Les motifs implémentés dans ce modèle sont inspirés des métriques d'interprétation du chapitre 4. Nous avons également développé un outil permettant de lire ce modèle de migration et de l'appliquer de manière automatisée à l'ensemble des notices de chaque jeu de données. Nous présentons plus de détails sur cet outil dans le chapitre suivant.

Nous avons donc appliqué un même modèle de migration sur les trois jeux de données. La figure 5.21 présente des résultats issus de cette expérimentation. Pour chaque jeu de données, la première colonne (Mapped fields) correspond au nombre de champs des notices qui sont considérés dans les règles de notre modèle. La deuxième colonne (Migrated data) présente la quantité de données du catalogue en entrée qui est présente dans le résultat de la migration. La troisième colonne (Applied patterns) décrit le taux des motifs, de notre modèle de migration, qui ont été activés par le processus. La quatrième colonne (Applied mappings) présente enfin le taux de correspondances, issues des différents motifs, qui ont été utilisées durant la migration.

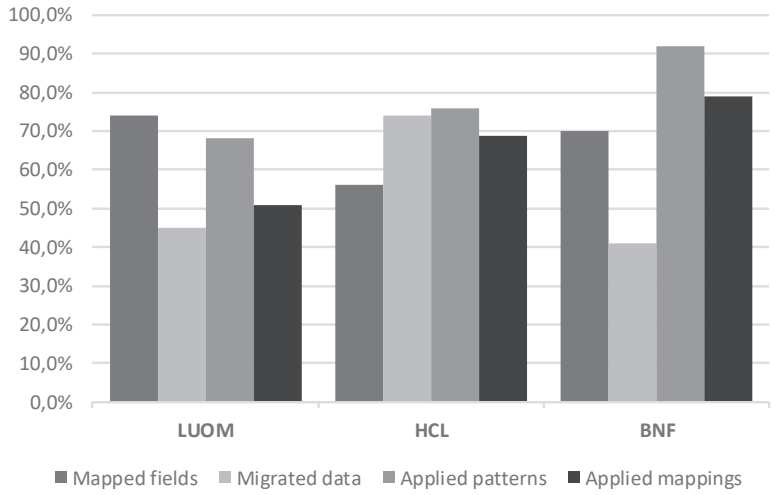


FIGURE 5.21 – Statistiques sur la détection et l'application d'un même modèle de règles de migration sur trois jeux de données différent

Dans les résultats de la figure 5.21, nous observons une différence entre le taux de champs détectés par l'outil (Mapped fields) et le taux effectif de données migrées (Migrated data), qui est inférieur d'environ 30%, pour les jeux de données LUOM et BNF. Cela veut dire que certains champs que nous ne considérons pas dans notre modèle décrivent une quantité importante de données dans le jeu de données correspondant. Pour ces deux jeux de données, il s'agit en réalité de champs spécifiques qui sont utilisés pour un usage interne, et qui concernent l'ensemble des notices, expliquant la quantité importante de données non-migrées. La considération de ces champs ne relève toutefois pas d'une interprétation des connaissances bibliographiques mais plutôt d'un paramétrage spécifique des données de l'institution. Les règles permettant l'intégration de ces données spécifiques ne peuvent pas être réutilisées et ne concernent donc pas directement notre étude. Les motifs (Applied Patterns) et leurs correspondances de migration (Applied Mappings)

5. `http://research.progilone.fr/patterns/COM3TMigrationModelRDF.xml`

étant axées sur les connaissances bibliographiques, nous obtenons un taux d'application des motifs de plus de 90% pour le jeu de données BNF car ce dernier (bibliothèque nationale) possède la plus importante richesse bibliographique. Pour les deux autres jeux de données, ce taux avoisine les 70%. Cela signifie qu'une majorité des motifs inclus dans notre modèle a donc été réutilisée dans les trois jeux de données de cette évaluation. Sur l'ensemble des jeux de données, au moins 50% des motifs et correspondances de migration peuvent être réutilisés dans différents contextes.

Ces observations nous indiquent qu'en possession d'un modèle de migration correctement structuré et orienté sur les motifs de connaissances bibliographiques, nous pouvons économiser certains efforts nécessaires à la configuration d'une solution dédiée à la migration d'un nouveau catalogue. En effet, cette expérimentation nous montre que les motifs de connaissances (*patterns*) peuvent avoir une forte capacité de réutilisation sur différents jeux de données hétérogènes. Nous admettons alors que l'effort de modélisation des motifs de connaissances (et d'appréhension de notre méthodologie) peut-être compensé par la capacité de réutilisation de ces derniers dans différents contextes. Dans le cas d'une institution ayant à réaliser plusieurs projets de migration de données bibliographiques, ce gain peut être crucial pour limiter les coûts de ces projets.

5.6 Conclusion

Nous avons détaillé notre méthode de méta-modélisation permettant la transformation de catalogues bibliographiques vers des bases de connaissances sémantiques. Notre approche est basée sur une instance d'un méta-modèle des connaissances bibliographiques qui s'applique aussi bien à des contextes bibliographiques élémentaires, avancés que spécifiques. L'avantage de sa structure en arbre de motifs réside dans sa capacité à réduire les problèmes de redondance dans l'écriture des correspondances de migration. De plus, la description des motifs à un niveau d'abstraction élevé facilite le travail des documentalistes et la réutilisation des motifs pour différents catalogues.

Du point de vue métier, notre contribution peut être adoptée de manière méthodologique. Cela veut dire que les documentalistes peuvent formuler les motifs de connaissances correspondants à leurs catalogues et les organiser en fonction des objectifs de leur projet de migration sans nécessiter de connaissances informatiques préalables. Ensuite, ces motifs abstraits sont traduits en un modèle de migration concret qui peut être implémenté dans un outil informatique. Se pose toutefois le problème des catalogues ayant des modèles de notices très spécifiques et difficiles à interpréter. La traduction des motifs en règles de migration et d'enrichissement peut être difficile à maintenir dans un contexte d'évolution des pratiques de catalogage. De plus la pérennité des correspondances d'enrichissement doit être assurée, si le modèle des sources externes évolue, tout en préservant la qualité et simplicité du modèle de règles. Nous admettons alors que l'adoption de nos principes de méta-modélisation n'est pas envisageable sans la réalisation d'une documentation détaillée sur les motifs qui sont crées et utilisés dans les outils de transformation.

Du point de vue technique, notre modèle de migration repose sur des mécanismes que seule une partie des outils de migration peut intégrer. Même si les outils s'améliorent avec le temps, il peut être très coûteux de développer une solution basée sur notre modèle pour un projet ponctuel, à la différence des outils utilisant par exemple XSLT qui peuvent convenir pour des catalogues simples. Par ailleurs, le développement des fonctions de transformations et conditions nécessaires à l'interprétation des notices peut également être un processus coûteux. Dans ce cas, il est en général nécessaire de réaliser différentes itérations des processus de migration et d'enrichissement pour adapter les règles progressivement. Ainsi, l'intérêt de notre méthodologie de modélisation reste dépendant de la possibilité de l'outil qui intègre cette méthode, de pouvoir facilement rejouer les processus de migration et d'enrichissement par les documentalistes et/ou informaticiens.

Chapitre 6

COM3ET, un outil de transformation de notices

Dans ce chapitre, nous présentons notre implémentation d'un système partiellement automatisé d'intégration de métadonnées bibliographiques. Notre système, reposant sur les concepts détaillés aux chapitres 4, 5 et 6, permet de transformer des catalogues de notices en bases de connaissances sémantiques. Le processus principal d'intégration des données s'articule autour de deux phases majeures qui sont (1) la construction et l'application d'un modèle de migration et d'enrichissement sur les notices et (2) la déduplication des entités produites. L'objectif de ce travail est, d'un côté, de montrer la faisabilité du processus d'intégration de notices selon notre méthodologie orientée motifs de connaissances et, d'un autre côté, d'explorer des méthodes permettant d'automatiser partiellement certaines phases du processus de transformation des notices.

Sommaire

6.1 Introduction

La gestion des métadonnées, dans la communauté documentaire, doit répondre à des exigences de qualité afin de respecter la richesse du patrimoine et les objectifs de valorisation des connaissances des institutions [51, 30]. Dans les chapitres précédents nous avons détaillé des méthodes d'intégration de données documentaires qui tiennent compte des spécificités des métadonnées bibliographiques. Cependant, la communauté a également soulevé la nécessité de réaliser de nouveaux outils automatisés pour la transformation des notices afin de permettre une migration

plus massive des catalogues existants vers les technologies du Web Sémantique [154, 54, 72]. La suite logique de nos travaux est donc de réaliser une implémentation du processus de transformation des notices en intégrant les principes de notre méthodologie orientée motifs de connaissances.

Dans ce chapitre, nous présentons le système **CoM3ET**, dédié à la migration et l'enrichissement des notices bibliographiques. Cet outil nous permet à la fois d'expérimenter de nouvelles méthodes d'interprétation et de transformation de notices, et d'étudier les possibilités offertes par notre méthodologie pour l'automatisation du processus global de transformation. Dans cette solution implémentée, le processus principal s'articule autour de deux phases successives qui sont (1) l'*application* d'un modèle de migration et d'enrichissement sur les données à transformer et (2) la *déduplication* des entités. Nous nous basons sur les métriques de qualité du chapitre 4, implémentées comme des fonctions, pour automatiser la création du modèle de migration et nous utilisons également une méthode originale pour automatiser la phase de *blocking* lors de la déduplication. Nous détaillons chacune de ces phases et nous illustrons nos propos avec des exemples d'utilisation relatifs à la transformation de notices de type MARC vers les principes FRBR/LRM.

Dans le reste de ce chapitre, la section 7.2 présente le périmètre d'implémentation de notre solution, c-à-d les restrictions liées aux données ainsi que les choix architecturaux associés. La section 7.3 décrit la phase de création et d'application du modèle de migration et d'enrichissement pour la création de la future base de connaissances et la section 7.4 détaille la phase de déduplication telle qu'elle est implémentée dans notre système.

6.2 Périmètre d'implémentation

Nous commençons par décrire les spécificités liées à l'implémentation de notre système appelé CoM3ET. Ce dernier, que nous détaillons ci-après, prend en entrée des notices bibliographiques, utilisant un format de type MARC comme MARC21 ou UNIMARC, et retourne une base de connaissances reposant sur les principes FRBR/LRM. Le processus de transformation des notices consiste notamment en une extraction des relations bibliographiques, représentées implicitement dans les notices, et d'une modélisation de ces dernières selon un graphe d'entités et de propriétés.

La figure 6.1 présente deux exemples de notices qui peuvent être placées en entrée de CoM3ET. Ces notices décrivent deux Œuvres qui sont *The Da Vinci Code* de *Dan Brown* et *O alquimista* de *Paulo Coelho*. Elles incluent toutes les deux des relations bibliographiques qu'il convient d'interpréter comme la *traduction* en français de l'Oeuvre *The Da Vinci Code* ou l'*augmentation* de l'Œuvre *O alquimista* avec des illustrations. Nous utilisons plus loin ces notices pour illustrer des exemples de fonctions d'interprétations qui sont intégrés dans notre solution.

6.2.1 Présentation du système CoM3ET

Nous appelons notre système **CoM3ET** pour Case-oriented MARC Metadata Migration & Enrichment Tool. Ce dernier prend en entrée trois éléments d'informations différents qui sont le catalogue de données à transformer, un ensemble de métriques et un méta-modèle de connaissances bibliographiques. Ces éléments ayant été détaillés respectivement dans les chapitres 4 et 5, nous les présentons ci-après d'un point de vue essentiellement fonctionnel. Notre système permet différents niveaux d'automatisation en fonction de son degré de configuration (notamment des métriques de qualité et des motifs de connaissances). Dans un cas idéal, où le système est entièrement configuré pour un catalogue donné, il est possible de réaliser toutes les étapes de la transformation des notices du catalogue de manière complètement automatique.

```
010 $a 2-7441-7554-4
101 $a fre $c eng
102 $a US
200 $a The Da Vinci code
    $f Dan Brown
210 $a Paris $d 2004
215 $a 523 p. $d 26 cm
700 $a Brown $b Dan $4 070
    $3 14529449
702 $a Roche $b Daniel $4 730
    $3 13329291
```

(a) The Da Vinci Code [a], UNIMARC

a. `bnf:ark:/12148/cb39929925k`

```
010 $a 972-711-011-8
101 $a por
102 $a PT
200 $a O alquimista $f Paulo Coelho
    $g rev. Maria Manuela Garcia Cruz
    $g ill. João Batel
210 $a Lisboa $c Pergaminho, $d 1995
215 $a 245, [4] p. $c il. $d 23 cm
700 $a Coelho, $b Paulo, $f 1947- $4 070
702 $a Cruz, $b Maria Manuela Garcia $4 640
702 $a Batel, $b João $4 440
```

(b) O alquimista [a], UNIMARC

a. `bnportugal:3100024~!528877~!0`

FIGURE 6.1 – Deux exemples de notices issues de deux bibliothèques nationales

Dans CoM3ET, la phase de migration et d'enrichissement des notices est décomposée en sous-étapes qui sont respectivement (1) la création du modèle de migration, (2) l'application de ce modèle pour créer les entités cibles puis (3) l'enrichissement de ces dernières avec des sources externes. La déduplication est une étape utilisée ponctuellement par les différentes sous-étapes (c-à-d., migration et enrichissement).

La figure 6.2 présente le schéma d'architecture de CoM3ET.

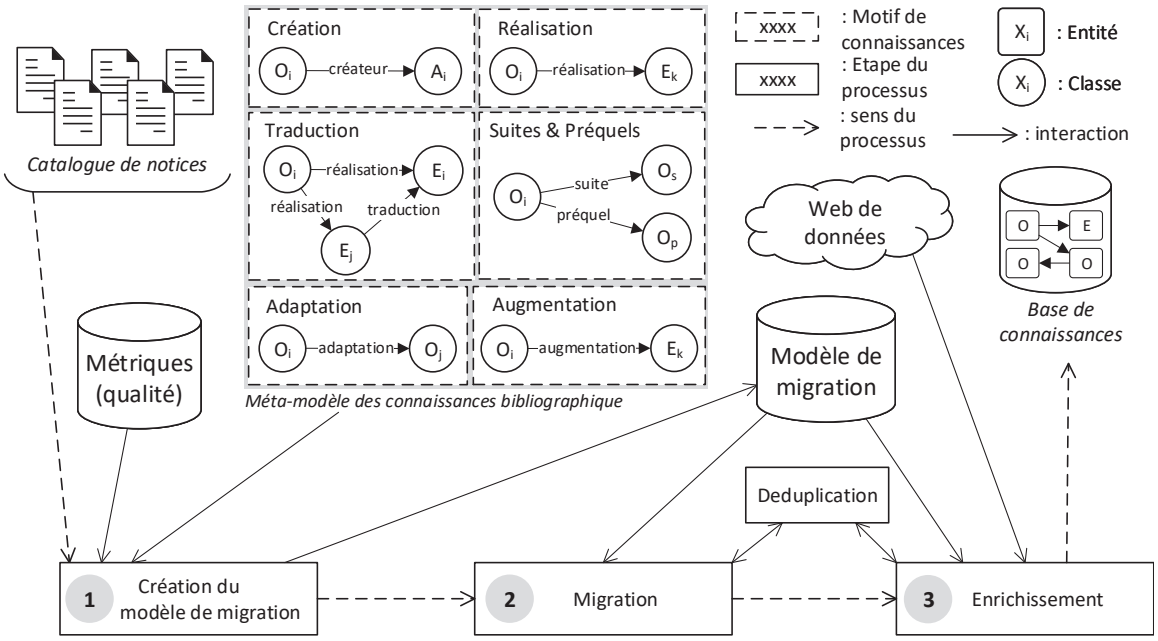


FIGURE 6.2 – Schéma d'architecture de CoM3ET

Sur la figure 6.2, le catalogue de notices initial (en haut à gauche) est progressivement transformé pour obtenir une base de connaissances (à droite). Les autres données en entrée sont des métriques de qualité, définies dans le chapitre 4, et un méta-modèle des connaissances bibliographiques comme détaillé dans le chapitre 5. Les métriques de qualité permettent d'interpréter les spécificités et erreurs de catalogage, ainsi que les connaissances bibliographiques avancées du

catalogue. Le méta-modèle sert à définir la manière dont les métadonnées interprétées doivent être représentées dans la future base de connaissances. Ces données sont utilisées à l'étape (1) pour construire le modèle de migration. Dans COM3ET, si toutes les métriques et tous les motifs, qui sont utiles à l'interprétation d'un catalogue donné, sont préalablement implémentées, alors le modèle de migration peut être créé automatiquement. Ce dernier sera ensuite utilisé aux étapes (2) et (3) pour traduire les notices en entités et propriétés, puis pour enrichir ces dernières avec des sources externes. Ces deux dernières étapes sont ponctuées de phases de déduplication permettant d'éliminer la redondance dans les données qui sont migrées ou enrichies.

6.3 Création et application du modèle de migration

Nous décrivons la partie de notre système permettant la création et l'application du modèle de migration. Sur la figure 6.2, il s'agit de la relation entre les sous-étapes (1) et (2). Un modèle de migration contient les règles permettant d'interpréter automatiquement les notices à migrer et de représenter les informations bibliographiques de ces notices dans une base de connaissances. Dans le contexte de notre solution CoM3ET, le modèle de migration est construit comme une instance spécifique du méta-modèle des connaissances bibliographiques, c'est à dire qu'il contient un ensemble de motifs où chaque motif modélise une des connaissances bibliographiques du catalogue à transformer (*cf.*, chapitre 5). Cette instance du méta-modèle n'utilise que les motifs qui sont pertinents pour la transformation d'un catalogue. De plus, nous pouvons ajouter des conditions sur chaque motif afin de restreindre leur application sur les notices.

Chaque motif décrit des correspondances entre les éléments qui le compose (c-à-d., entités, propriétés) et les informations décrites dans les notices. Nous parlons "d'informations" car les données brutes, issues des champs des notices doivent être préalablement filtrées, nettoyées ou combinées par des fonctions spécifiques avant d'être intégrées dans un motif. Ces fonctions sont donc associées aux correspondances dans les motifs. La figure 6.3 présente un schéma qui illustre les relations entre le modèle de migration de CoM3ET et les données des notices à transformer.

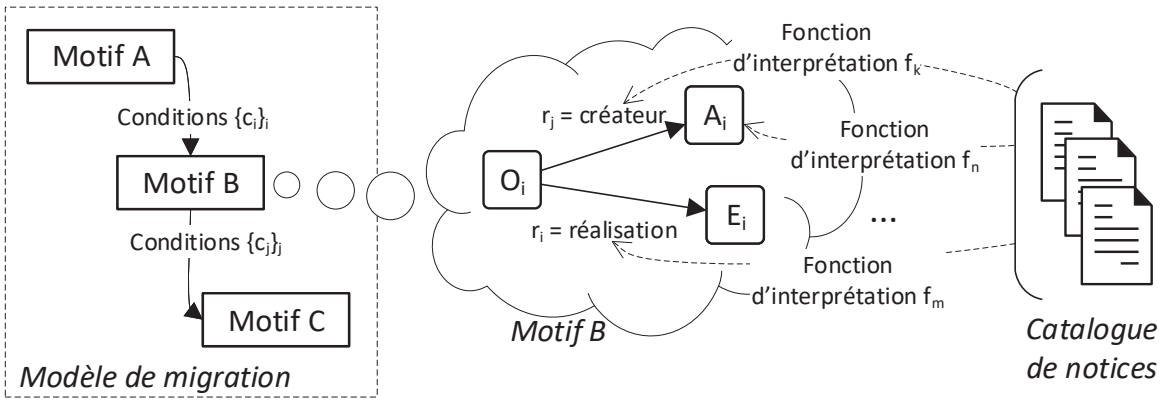


FIGURE 6.3 – Lien entre le modèle de migration et les notices du catalogue

Dans cet exemple, figure 6.3, le motif B décrit trois entités FRBR, O_i (*Œuvre*), A_i (*Agent*) et E_i (*Expression*). Lors du traitement automatique d'un catalogue, si le motif A a été appliqué sur une notice donnée, le système évalue l'ensemble des conditions $\{c_i\}i$ sur cette même notice. Si les conditions sont validées, alors le système va créer les entités et relations associées au motif B selon le résultat des fonctions d'interprétations sur la notice. Par exemple, la relation de création entre l'Œuvre O_i et l'Agent A_i dépend de la fonction d'interprétation f_k. Cette dernière peut par exemple retourner une valeur booléenne selon le code de fonction de l'agent dans la notice.

Sur la notice de la figure 6.1, la fonction d'interprétation f_k doit identifier l'Agent *Paulo Coehlo* comme auteur (code 070 du champ 700$4) de l'Œuvre *O alquimista*.

Création des fonctions d'interprétation

Nous décrivons maintenant la construction des fonctions d'interprétation dans notre système et leur intégration dans les motifs qui doivent être utilisés pour la transformation d'un catalogue. Pour rappel, ces fonctions d'interprétations retournent une information utile sur une notice afin de déclencher la création d'une entité ou d'une propriété d'un motif de connaissances. Les fonctions d'interprétation consistent en une description logique de l'extraction des données des notices. La conception de ces fonctions ne peut être automatisée qu'à la condition que les spécificités du format des notices en entrée soient connues. Dans CoM3ET, les fonctions d'interprétation sont basées sur les métriques de qualité du chapitre 4 et produisent, lors de l'évaluation des champs des notices, une valeur booléenne ou textuelle à exploiter. Par exemple, nous utilisons la métrique MD (données manquantes) pour créer des fonctions comme *MD-langue* ou *MD-résumé* nécessaires à la détection et l'interprétation de certains motifs comme la traduction ou l'adaptation d'Œuvres. L'écriture de ces fonctions repose, dans COM3ET, sur une collecte préalable (et partiellement automatisée) des statistiques sur les champs qui sont utilisés dans le catalogue.

Nous présentons un exemple pratique de fonction d'interprétation. Parmi les motifs à considérer dans la figure 6.2, le motif de *création* définit l'Œuvre principale ainsi que sa relation avec un ou plusieurs Agents créateurs. La définition de l'Œuvre implique, par exemple, de lui attribuer une propriété *titre*, issue des données de la notice à transformer. Nous proposons une fonction permettant de créer ce titre pour l'Œuvre. Son principe consiste à utiliser la métrique MD pour vérifier, au préalable, que la notice traitée possède au moins un *titre de publication*, puis, en cas de titres multiples, d'utiliser la métrique MUT (présence d'un titre uniforme) pour définir la provenance du titre de l'Œuvre. Si un *titre uniforme* est détecté par cette métrique, il doit être utilisé pour la valeur du titre de l'Œuvre, sinon c'est le titre de publication qui doit être utilisé.

Exemple 6.3.1. Interprétation du titre de l'Œuvre. Nous proposons de formuler cette fonction avec une notation en logique du premier ordre : $(\neg MD('\texttt{t\_pub}')) \wedge ((\neg MUT('\texttt{t\_uni}') \wedge \sigma_{titreOeuvre} = val('\texttt{t\_uni}')) \vee (MUT('\texttt{t\_uni}') \wedge \sigma_{titreOeuvre} = val(('\texttt{t\_pub}'))))$ avec t_pub le titre de publication [1], t_uni le titre uniforme [2], MD la métrique déterminant si un champ est manquant, MUT déterminant si le titre uniforme est manquant et val retournant la valeur d'un champ. $\sigma_{titreOeuvre}$ correspond à la valeur obtenue par l'application de cette fonction.

Ce exemple nous montre que les fonctions d'interprétation peuvent être utilisées pour appliquer des conditions sur les champs des notices et/ou récupérer des valeurs dans ces derniers. Les fonctions peuvent également transformer les données avant de les intégrer (cf., exemple suivant).

Exemple 6.3.2. Ajout d'un préfixe sur un identifiant. Les notices décrivant des ouvrages contiennent en général un champ pour l'identifiant ISBN [3] de l'ouvrage. Il peut être nécessaire de reprendre cet identifiant lors d'un processus de migration. Cependant, si le modèle distant ne propose pas de propriété explicite pour cet ISBN, une solution peut consister à utiliser une propriété à la sémantique plus large, comme `rda:identifierForTheManifestation` avec FRBR, et d'ajouter un préfixe "ISBN :". Pour la notice de la figure 6.1a, cela donne la valeur "ISBN :2-7441-7554-4" que l'on ajoute à une entité de classe FRBR Manifestation.

COM3ET intègre plusieurs dizaines de fonctions que nous répartissons en deux catégories. D'un côté, des fonctions permettent de transformer les valeurs des notices comme par exemple en

1. `http://bnf.fr/documents/B200_6_2011.pdf`
2. `http://bnf.fr/documents/B500_6_2010.pdf`
3. `http://bnf.fr/fr/professionnels/s_informer_obtenir_isbn/s.qu_est_ce_que_isbn.html`

ajoutant un préfixe à une valeur migrée ou en remplaçant un code dans une notice par une valeur issue d'une table de correspondance. Ces fonctions sont associées aux correspondances de migration dans les motifs de connaissances. D'un autre côté, d'autres fonctions de COM3ET sont destinées à l'application des métriques de qualité sur les notices. Par exemple, la métrique RLE (les responsabilités n'ayant pas d'identifiants) déclenche la création ou la récupération d'identifiants à la volée pour les entités de classe Agent. La métrique CPN (pratiques de catalogage) peut indiquer la présence de données codées, nécessitant alors l'intégration de données externes pour les décoder. Ces fonctions sont notamment utilisées pour analyser un catalogue en entrée et collecter des statistiques sur ses spécificités mais aussi pour créer des conditions, associées aux motifs ou à leurs correspondances afin de correctement interpréter les connaissances d'un catalogue. La figure 6.4 présente une capture d'écran de COM3ET affichant des statistiques sur un catalogue analysé par des fonctions d'interprétation.

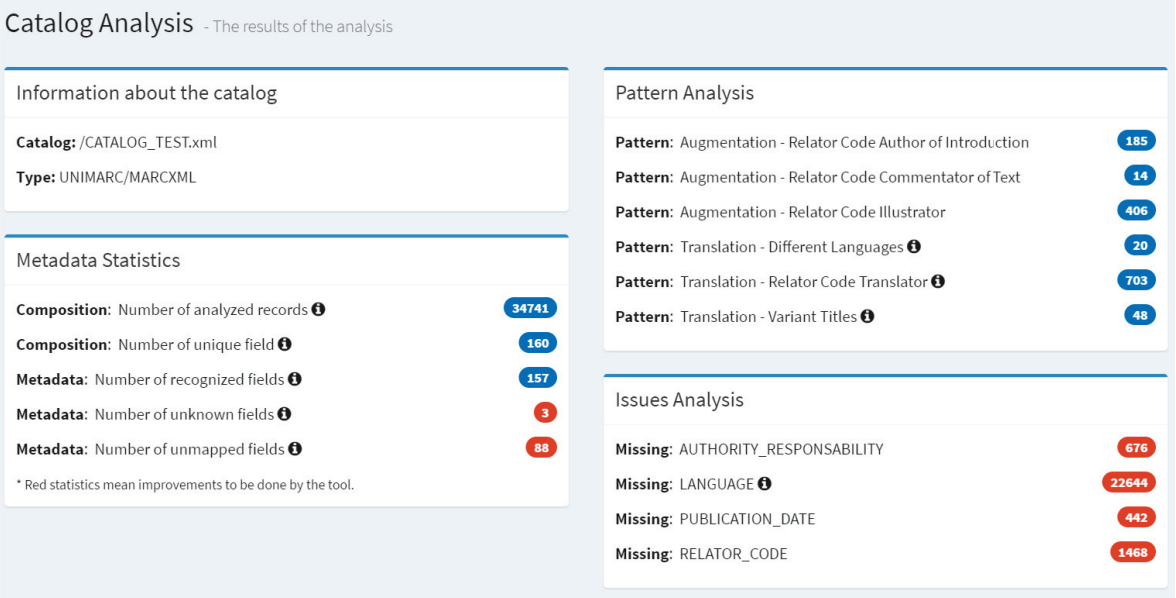


FIGURE 6.4 – Interface d'analyse d'un catalogue dans COM3ET

Nous pouvons observer dans cette capture d'écran (*cf.*, figure 6.4) que ce catalogue d'environ 35.000 notices peut contenir des relations bibliographiques d'augmentation et de traduction, et qu'il manque la langue de publication des ressources dans une grande majorité des notices. De plus, seule la moitié des champs utilisés sont considérés par des correspondances existantes dans les motifs du modèle de migration. Ces informations permettent notamment d'orienter les efforts nécessaires à la réalisation du modèle final de migration et d'enrichissement.

Instanciation des motifs de connaissances

La construction du modèle de migration s'effectue en principe une fois que toutes les fonctions d'interprétation sont définies et associées aux motifs de connaissances. Pour rappel, ces fonctions sont déduites des métriques de qualité détaillées au chapitre 4. La dernière étape consiste donc à organiser les différents motifs de connaissances en un arbre qui sera parcouru pour chaque notice à transformer. Pour cela, nous appliquons les principes détaillés dans le chapitre 5.

Dans COM3ET, nous avons préalablement défini plusieurs dizaines de motifs communs aux catalogues bibliographiques. Ce travail a consisté essentiellement à modéliser manuellement les familles bibliographiques de la littérature, [137, 133], avec les classes et propriétés FRBR et

RDA, en suivant les recommandations du domaine [32, 114]. Lors du traitement d'un catalogue contenant des spécificités, nous n'avons qu'à intégrer des motifs spécifiques pour compléter notre modèle. De plus, nous automatisons le chargement du modèle dans le système. En effet, lors de l'analyse d'un catalogue, nous comparons préalablement les champs que nous détectons dans les notices avec les conditions des motifs pré-existants pour sélectionner ceux qui seront pertinents pour construire *à la volée* le modèle de migration. La figure 6.5 présente une capture d'écran de COM3ET montrant un modèle de migration chargé automatiquement à la lecture d'un catalogue.

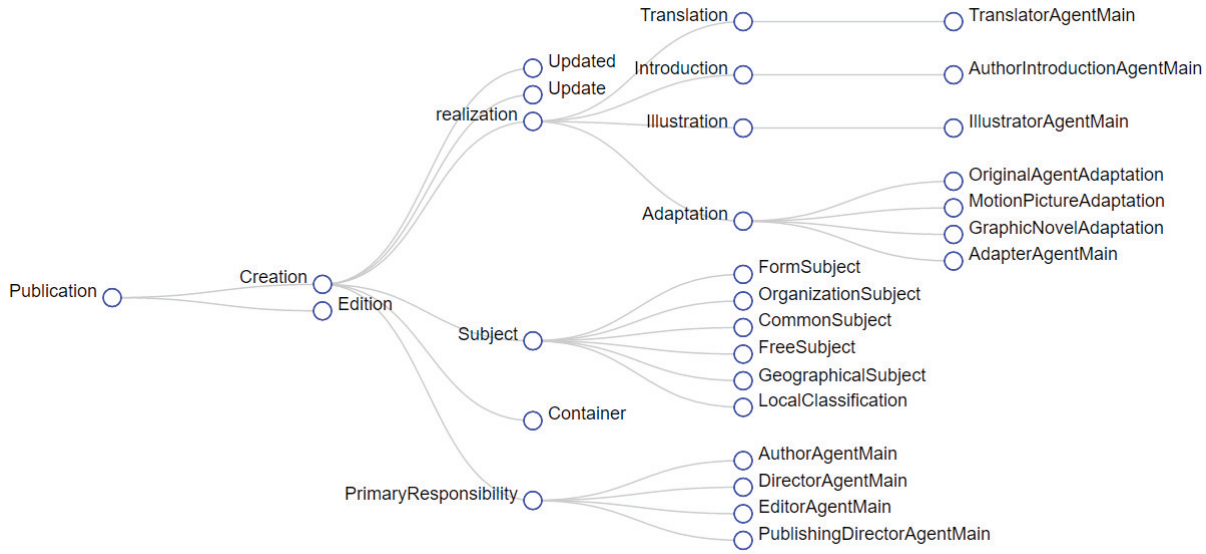


FIGURE 6.5 – Représentation d'un modèle de migration dans COM3ET

Dans cet exemple, le motif racine de l'arbre est le motif *Publication* qui déterminera si la notice évaluée contient *a minima* les informations basiques de description de la publication d'un document (*ex.*, titre, identifiant). Si ce motif est validé, la notice sera traitée en appliquant en premier lieu le motif *Publication*. Les deux motifs suivants, *Création* et *Édition*, relèvent respectivement de la détection et modélisation de l'Œuvre originale de cette publication et de la détection de potentielles éditions multiples de cette publication (*ex.*, éditions électroniques, rééditions). Les motifs suivant sont essentiellement des spécialisations de ces motifs. Cette organisation des motifs dépend des conditions qui leur sont associées. En effet, nous choisissons de bénéficier de l'héritage entre les motifs d'une même branche de l'arbre et organisons les conditions des moins restrictives, vers la racine, aux plus restrictives, vers les feuilles. Dans l'exemple de la figure 6.5, le motif *TranslatorAgentMain* relatif au traducteur, hérite des conditions sur le motif *Translation*. De plus, nous utilisons des motifs de haut niveau pour décrire les entités principales du modèle cible. Ainsi, dans cet exemple, les attributs d'un Agent (ex., nom, prénom) décrits dans le motif *PrimaryResponsability* sont hérités par les motifs *AuthorAgentMain*, *DirectorAgentMain* etc.

Lorsque le modèle est construit, ou chargé automatiquement, il peut être appliqué aux notices du catalogue. Là, l'arbre des motifs est parcouru sur chaque notice et les entités et propriétés des motifs valides sont progressivement construites. Comme détaillé au chapitre 5, nous appliquons un mécanisme de contextualisation des classes, décrites dans les motifs, lors du traitement d'une notice afin d'associer correctement les propriétés aux entités correspondantes quand une même classe (d'entité) est réutilisée dans plusieurs motifs. Les entités produites sont ensuite traitées par le processus de déduplication, que nous détaillons plus bas dans ce chapitre.

Une fois que les entités sont migrées et dédupliquées, l'utilisateur de COM3ET peut sélectionner une entité dont il souhaite enrichir les informations. Dans ce cas, notre système parcourt à nouveau le modèle de migration à la recherche de motifs possédant des correspondances vers des sources externes. Par exemple, COM3ET intègre des correspondances d'enrichissement sur les motifs d'adaptation, de réalisations, de sequels et préquels ou encore d'agrégations. Ces correspondances sont décrites dans le chapitre 5. Lorsqu'une correspondance d'enrichissement est appliquée, une requête *SPARQL* est déclenchée sur la source de données en question et le résultat est dédupliqué avec les entités locales puis intégré à l'entité préalablement sélectionnée. Cette intégration est réalisée grâce aux correspondances d'enrichissement qui permettent les alignements entre les propriétés et les classes des modèles. En cas de doublons, nous soumettons les équivalences à l'utilisateur qui se charge de ne conserver que la donnée qui l'intéresse. Sur la figure 6.6, nous reprenons l'exemple de la notice décrivant l'Œuvre *The Da Vinci Code* et nous présentons une capture d'écran d'un résultat proposé par COM3ET pour l'enrichissement de cette Œuvre.

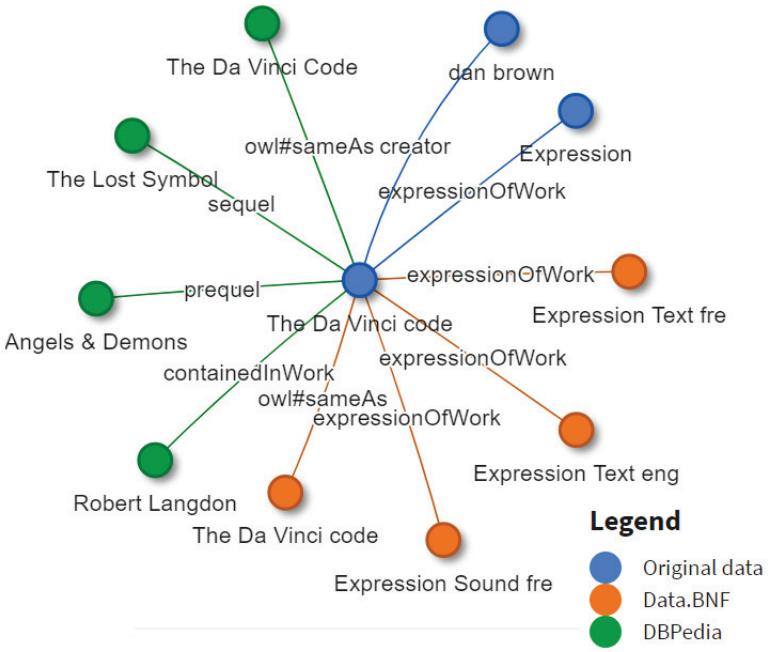


FIGURE 6.6 – Résultat visuel d'un processus d'enrichissement dans COM3ET

Dans cet exemple, deux sources de données ont été utilisées pour enrichir l'Œuvre *The Da Vinci Code*, *data.BNF* et *DBPedia*. Les données locales (Original Data) consistent en la description de l'Œuvre (au centre) et de deux propriétés, une décrivant l'Expression locale (le texte en français), l'autre décrivant l'auteur *Dan Brown*. Un motif sur les réalisations d'une Œuvre a permis d'extraire de nouvelles Expressions depuis *data.BNF* comme le texte en anglais ou une version audio en français (*Expression Sound Fre*). Aucune Expression n'est extraite de *DBPedia* car ce niveau d'information n'existe pas dans cette source. En revanche, le motif concernant les préquels et les suites permet d'extraire les entités correspondantes *(Angels & Demons)* et (*The Lost Symbol*) depuis *DBPedia*. Comme la structure des informations peut varier entre le modèle local (*ex.*, FRBR) et le modèle distant (*ex.*, DBPedia ontology), nous avons associé des fonctions spécifiques aux opérateurs utilisés dans les correspondances d'enrichissement afin de moduler les requêtes SPARQL comme l'opérateur (^) impliquant l'utilisation d'une relation inverse dans la requête (*ex.*, *contient* ou *estContenuPar*). Dans la section 6.5 de ce chapitre, nous présentons plus en détail les mécanismes de COM3ET pour l'enrichissement sémantique des entités.

6.4 Module de déduplication automatisé

Nous avons vu au chapitre 3 que le caractère automatique d'un processus de déduplication est limité par la nécessité de définir des clés de blocking pour le traitement de grands volumes d'entités. Dans COM3ET un module de déduplication automatique a été développé intégrant les phases de *blocking* et de *matching*. L'étape de blocking est elle-même automatisée grâce à un procédé de suggestion automatique de clés. L'objectif est d'assister l'expert dans le paramétrage de la déduplication et, en même temps, de proposer une méthode rapide pour réaliser le processus de déduplication sans interruption manuelle.

6.4.1 Détection des attributs candidats au blocking

Nous rappelons que le principe du blocking consiste à créer des blocs d'entités à comparer selon certains critères, limitant le nombre total de comparaisons à effectuer et permettant le calcul parallèle des blocs. Ces critères s'incarnent principalement sous la forme de clés issues d'une combinaison des attributs des entités. Si deux entités d'une même classe possèdent un attribut ayant une sémantique identique (*ex.*, date de publication), l'attribut en question peut être utilisé comme clé de blocking. Ce type de clé doit servir à former les groupes dans lesquels les algorithmes de déduplication vont comparer les entités. C'est pourquoi, les clés de blocking sont construites à partir des attributs qui ne sont pas utilisés pour la comparaison des entités. Dans le contexte des métadonnées bibliographiques, plusieurs attributs ou combinaisons d'attributs peuvent être des candidats à la création d'une clé de blocking comme l'*année de création* d'une Œuvre ou le *prénom de l'auteur* principal. Cependant, certaines combinaisons de propriétés peuvent être plus pertinentes que d'autres par leur capacité à maximiser la taille et l'homogénéité des groupes d'entités obtenus. En effet, il n'est pas efficace d'utiliser une clé qui soit trop discriminante, c'est à dire qui entrainerait la création d'une majorité de groupes de seulement une entité.

COM3ET intègre un processus automatique de suggestion de clés de blocking qui soient susceptibles de former des blocs de tailles équilibrées. Ce processus repose sur la création d'un treillis des combinaisons d'attributs potentiels. La Figure 6.7 illustre ce procédé avec un exemple de treillis pour déterminer la clé de blocking d'une œuvre bibliographique selon quatre attributs, l'*année de création*, le *nom de(s) l'auteur(s)*, le(s) *sujet(s) de l'œuvre* et le *genre de l'œuvre*.

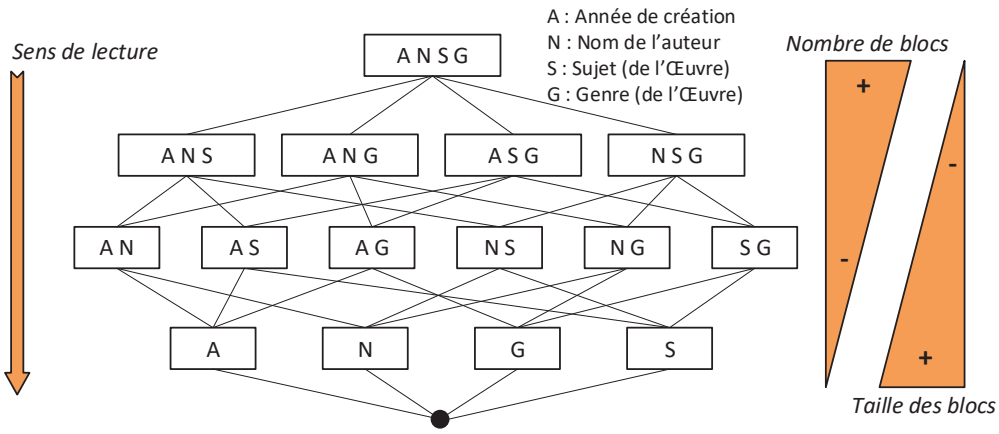


FIGURE 6.7 – treillis d'analyse de la meilleur clé de blocking

Chaque niveau du treillis équivaut à un rang, c'est à dire un nombre fixé d'attributs qui sont combinés. La racine représente donc chaque attribut seul comme clé de blocking. Le résultat de chaque nœud du treillis consiste en des statistiques sur la taille des blocs formés en groupant

les entités selon l'équivalence de la clé de blocking obtenue par la combinaison proposée. Par exemple, au premier rang, chaque entité se voit attribuer une clé de blocking calculée à partir des quatre attributs sélectionnés (ANSG), puis les entités sont groupées selon les équivalences de leurs clés respectives. La recherche de la meilleurs combinaison consiste alors à parcourir le treillis de haut en bas et d'évaluer, à chaque niveau, quelle combinaison est la plus adaptée à la création de blocs d'entités. De manière pratique, cela revient à calculer la moyenne harmonique des tailles de blocs créés pour chaque combinaison et ne conserver que la ou les combinaisons dont la moyenne est la plus élevée. Par exemple, si la combinaison SG pour *Sujet* et *Genre* obtient la moyenne harmonique la plus élevée, elle sera retenue comme combinaison pour la clé de blocking.

Cette méthode n'exclue pas la validation de l'expert des clés. Néanmoins, elle donne une indication à ce dernier et lui facilite l'analyse des entités pour prendre sa décision.

6.4.2 Évaluation de la similarité des pairs d'entités

L'évaluation de la similarité de deux entités dépend d'une ou plusieurs clés permettant la comparaison des entités et de la fonction de similarité appliquée sur ce(s) clé(s). Dans le contexte bibliographique, la génération de ces clés repose essentiellement sur les travaux de l'OCLC [4]. Mitchell et McCallum, dans [91], ont réalisé plusieurs expérimentations permettant de valider l'efficacité de ces clés. Dans COM3ET, nous réutilisons les résultats de ces travaux pour la création des clés de comparaison des paires présentes dans un bloc.

Pour évaluer la similarité d'une Œuvre, nous considérons une séquence ordonnée de patrons de clés qui sont appliqués en fonction des données disponibles (*ex.*, les données qui ont été migrées des notices vers les entités à comparer). La liste suivante présente ces patrons dans l'ordre d'importance avec leur niveau de pertinence :

— < Titre uniforme > (Forte pertinence)

— < Traducteur - Titre original > (Forte pertinence)

— < Nom du créateur - Titre de publication > (Forte pertinence)

— < Nom de la première responsabilité - Titre de publication > (Moyenne pertinence)

— < Nom de la première responsabilité - Premier titre disponible > (Faible pertinence)

Dans COM3ET, nous avons également intégré le processus PairRang [74] (cf., chapitre 3) permettant de forcer l'équilibrage des blocs si le processus de blocking ne le permet pas en amont. Une fois que cet équilibrage est effectué, COM3ET compare les entités au sein d'un bloc en utilisant les clés de *matching* qu'il évalue selon une combinaison de mesures de similarités sur les chaînes de caractères. Cette combinaison est définie manuellement selon le projet de migration ou d'enrichissement à effectuer. Cependant, la configuration par défaut de COM3ET utilise la mesure de Jaro adaptée par Winkler [146]. Nous faisons ce choix en nous basant sur les résultats de cette étude [66] et sur le fait que cette version de Jaro gère les légères variations typographiques entre les chaînes (*ex.*, Peter J et Peter John comme prénom), ce qui est pertinent vis à vis des notices dont le catalogage, réalisé manuellement, est sujet aux erreurs de saisies.

4. `https://www.oclc.org/content/dam/research/activities/frbralgorithm/2009-08.pdf`

6.5 Extraction de nouveaux motifs bibliographiques

Nous présentons les caractéristiques du processus d'extraction de motifs de connaissances dans des sources de données externes avec COM3ET.

6.5.1 Vue d'ensemble du processus

Le processus d'enrichissement sémantique correspond à l'ensemble des actions permettant d'augmenter la compréhension de données en entrée du processus. Dans notre contexte, le processus prend en entrée une entité sélectionnée (*ex.*, de classe *Œuvre* ou *Agent*), issue d'une base de données locale, et intègre de nouvelles connaissances bibliographiques à cette entité à partir de sources de données externes. Ces connaissances correspondent aux motifs bibliographiques que nous avons décrits au chapitre 5. Les informations locales sur l'entité à enrichir sont supposées déjà structurées car issues de la migration de notices bibliographiques selon notre méthodologie. Les connaissances externes doivent être caractérisées selon les motifs bibliographiques afin de permettre leur intégration. La Figure 6.8 présente un exemple de processus d'enrichissement sémantique d'une entité de classe *Œuvre* selon trois motifs de connaissances bibliographiques.

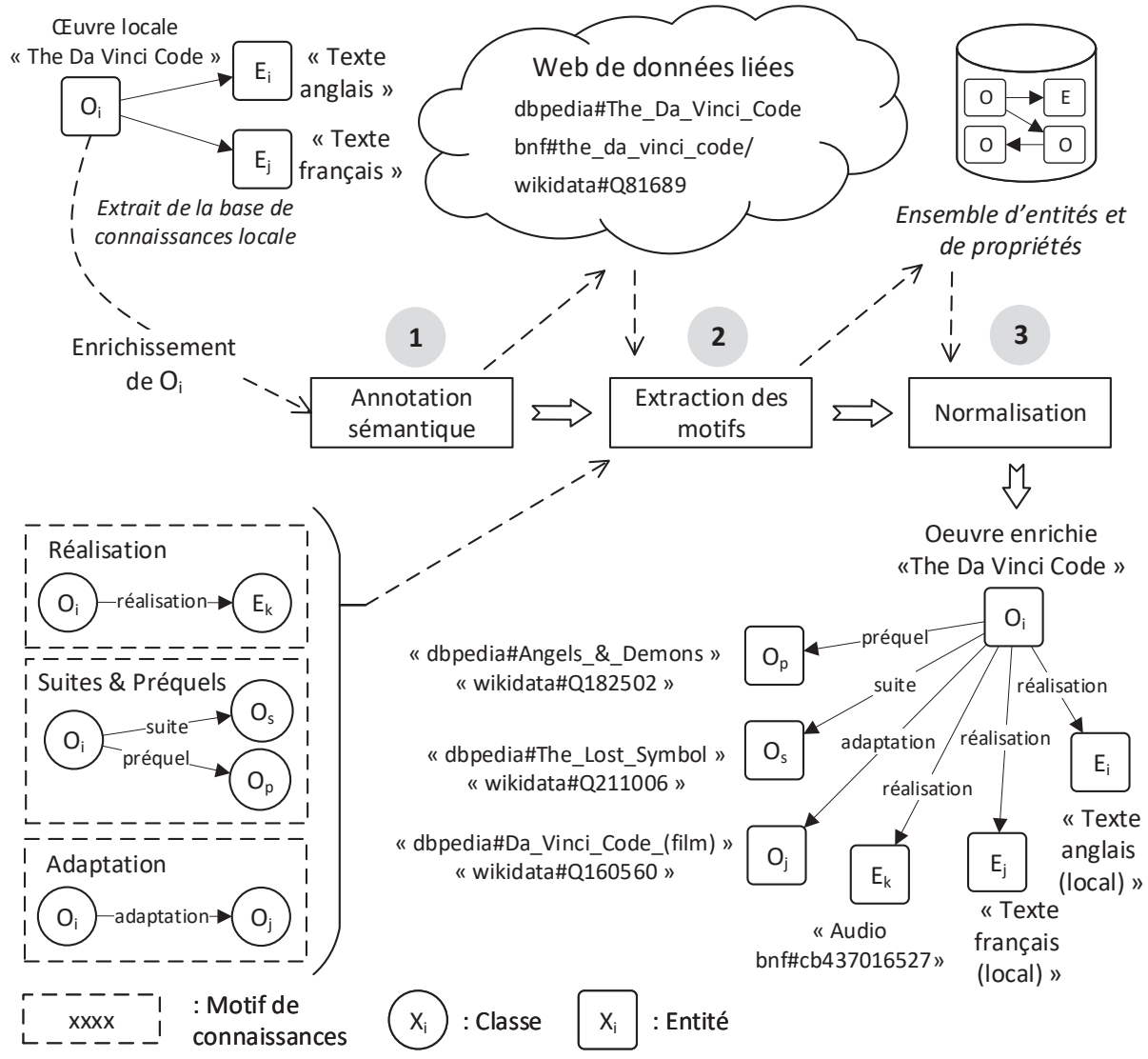


FIGURE 6.8 – Vue d'ensemble du processus d'enrichissement sémantique

Dans la figure 6.8, un processus d'enrichissement sémantique est appliqué à l'Œuvre O_i, *The Da Vinci Code*, sélectionnée dans une base de données locale. L'enrichissement permet ici d'intégrer de nouvelles informations selon trois motifs de connaissances, *Réalisation*, *Suites & Préquels* et *Adaptation*. A l'issue du processus, l'Œuvre est enrichie de nouvelles informations concernant, un préquel O_p (*Angels & Demons*), une suite O_s (*The Lost Symbol*), une adaptation O_j (Film éponyme) et une (nouvelle) réalisation E_k (*Audio*).

Le processus d'enrichissement s'effectue selon trois étapes, numérotées de 1 à 3. La première étape, **Annotation sémantique**, consiste à aligner une entité locale avec les entités, issues des sources de données distantes, qui sont équivalentes. Cette tâche, aussi appelée "Entity Linking", a été la source de multiples travaux et approches dans la communauté [128, 109, 50, 117]. Dans notre exemple, cette étape associe des relations de type *sameAs* à l'entité locale (Œuvre) O_i et des entités distantes (*ex.*, *wikidata#Q81689*). La deuxième étape, **Extraction des motifs**, prend en entrée l'entité locale O_i, complétée de liens d'équivalences, ainsi que les motifs de connaissances (en bas à droite de la figure). A partir de correspondances entre les classes et propriétés de chaque motif et les modèles des sources distantes, un ensemble d'entités et de propriétés, issues du web de données, sont extraits et liés à l'entité locale O_i. La dernière étape, **Normalisation**, contient l'ensemble des sous-tâches nécessaires au processus pour retourner un graphe de connaissances cohérent autour de l'entité locale enrichie. La plus importe de ces sous-tâches est la déduplication des entités équivalentes qui peuvent être extraites de plusieurs sources différentes. Dans notre exemple, l'Œuvre locale O_i est adaptée en un film qui est présent à la fois dans DBPedia (*Da_ Vinci_ Code_ (film)*) et dans Wikidata (*Q160560*).

6.5.2 Motifs implicites dans les sources externes

COM3ET permet d'extraire des connaissances depuis différentes sources externes en considérant que ces même connaissances sont représentées implicitement dans ces sources. Nous avons décrit, dans le chapitre 5, les différents mécanismes de notre méta-modélisation des règles afin de permettre cette extraction de connaissances implicites. Dans cette partie, nous montrons le contexte d'application de ces mécanismes dans notre implémentation de la solution COM3ET.

Pour rappel, nous considérons que la structure des connaissances à enrichir est préalablement définie par les motifs de connaissances bibliographiques. L'exemple de la figure 6.8 décrit trois motifs qui sont la *Réalisation*, les *Suites & Préquels* et l'*Adaptation* d'une entité, ici de classe Œuvre. Pour chaque motif, les sources distantes peuvent soit, ne pas contenir une telle information, la décrire de manière explicite (*ex.*, triplets correspondants aux propriétés du motif) ou encore la contenir mais de manière implicite. COM3ET, grâce à la structure des correspondances d'enrichissement (*cf.*, chapitre 5) permet de décrire le *chemin* dans la source externe pour extraire une information implicite.

Les informations implicites, dans notre contexte, sont des données nécessitant un effort d'interprétation pour être intégrée dans un processus d'enrichissement d'entités. Aalberg *et al*, dans [5], ont mis en évidence les problèmes d'interprétation des relations bibliographiques dans le cas de la migration des notices vers un modèles sémantique. Dans le contexte de l'enrichissement, le travail d'interprétation des sources distantes est similaire à celui dédié aux notices bibliographiques. A la différence des données implicites, qui peuvent être interrogées directement, l'interprétation des informations qui sont implicitement représentées dans une source nécessite d'extraire un ensemble plus large d'informations afin d'effectuer un traitement spécifique sur ces données.

Afin d'illustrer cette problématique, nous présentons quatre sources de données contenant des informations bibliographiques qui sont utilisées par COM3ET, *data.BNF*, *Worldcat*, *DBPedia* et *Wikidata*. Ces quatre sources intègrent les principes du web de données et sont de plus en plus utilisées dans le cadre de l'enrichissement de données bibliographiques. La table 6.1 compare trois niveaux de prise en charge des quatre sources pour sept motifs de connaissances bibliographiques. *Oui* signifie que l'information (c'est à dire le motif de connaissance) est contenue de manière explicite dans la source. *Non* signifie que l'information n'est pas représentée dans la source. *Analyse* indique que l'information peut être contenue dans la source, uniquement de manière implicite, et nécessite un travail d'interprétation pour déduire cette information.

	data.BNF	**Worldcat**	**DBPedia**	**Wikidata**
Création	Oui	Oui	Oui	Oui
Réalisations	Oui	Analyse	Non	Non
Traductions	Oui	Analyse	Analyse	Analyse
Suites	Non	Non	Oui	Oui
Adaptations	Analyse	Non	Analyse	Oui
Agrégations	Non	Non	Oui	Oui
Sujet Œuvre	Non	Analyse	Analyse	Analyse

TABLE 6.1 – Représentation de 7 motifs de connaissances dans 4 sources de données

Les quatre sources intègrent la notion de *Création* c'est à dire qu'il est possible d'extraire, a minima, une Œuvre bibliographique et ses créateurs. En effet, *data.BNF* et *Worldcat* intègrent les bases des principes FRBR/LRM et *DBPedia* et *Wikipedia* décrivent uniquement des œuvres notables. Dans le reste du tableau, on distingue la séparation entre les sources de données construites essentiellement à partir de données bibliographiques, *data.BNF* et *Worldcat*, et les sources plus généralistes à savoir *DBPedia* et *Wikidata*. En effet, les motifs de *Réalisations* et *Traductions* sont propres à la description documentaire, que l'on ne retrouve pas nécessairement dans les bases de données encyclopédiques. Ces deux motifs impliquent une représentation des Expressions intellectuelle d'une Œuvre. Les *Traductions*, seulement explicites dans data.BNF peuvent toutefois être déduites ou approximées dans les autres sources si une analyse préalable est effectuée sur les langues disponibles d'une même Œuvre. Dans le reste de cet exemple, on constate que les relations entre les Œuvres, nécessaires à la description des motifs de connaissances *Suites*, *Adaptations* ou *Agrégations* sont assez mal représentées dans les sources de données bibliographiques et qu'il est nécessaire d'aller les extraire dans des sources encyclopédiques. Enfin, le motif *Sujet Œuvre* décrit une relation entre deux Œuvres où l'une est sujet de l'autre. Cette information précieuse n'est pas disponible de manière explicite, pourtant très intéressante dans une famille bibliographique, et nécessite, quand c'est possible, d'analyser les sujets d'une Œuvre pour détecter d'autres œuvres potentielles.

6.6 Validations expérimentales

Dans cette section, nous présentons des expérimentations réalisées avec COM3ET dans le contexte industriel du projet Syrtis de l'entreprise Progilone. L'objectif de ces expérimentations est de valider les bénéfices de l'implémentation des métriques d'interprétation des connaissances bibliographiques (*cf.*, chapitre 4) couplée à notre méthode de méta-modélisation (*cf.*, chapitre 5) dans le cadre de ces projets de FRBRisation concrets.

Configuration industrielle

Pour réaliser ces expérimentations sur des catalogues réels, une version de COM3ET a été développée dans un cadre industriel. Cette version est associée au système d'information documentaire Syrtis, développé par la société Progilone [5]. Pour rappel, Syrtis permet notamment d'intégrer, de gérer et de diffuser des données basées sur les formats FRBR/LRM et plus généralement sur les données utilisant les principes du web sémantique. L'outil de migration COM3ET est connecté à Syrtis par un transfert automatique des métadonnées migrées au format RDF. La figure 6.9 présente une interface de Syrtis contenant la description d'une Œuvre selon le format FRBR, c'est à dire les quatre niveaux de classes *Œuvre*, *Expression*, *Manifestation* et *Exemplaire*.

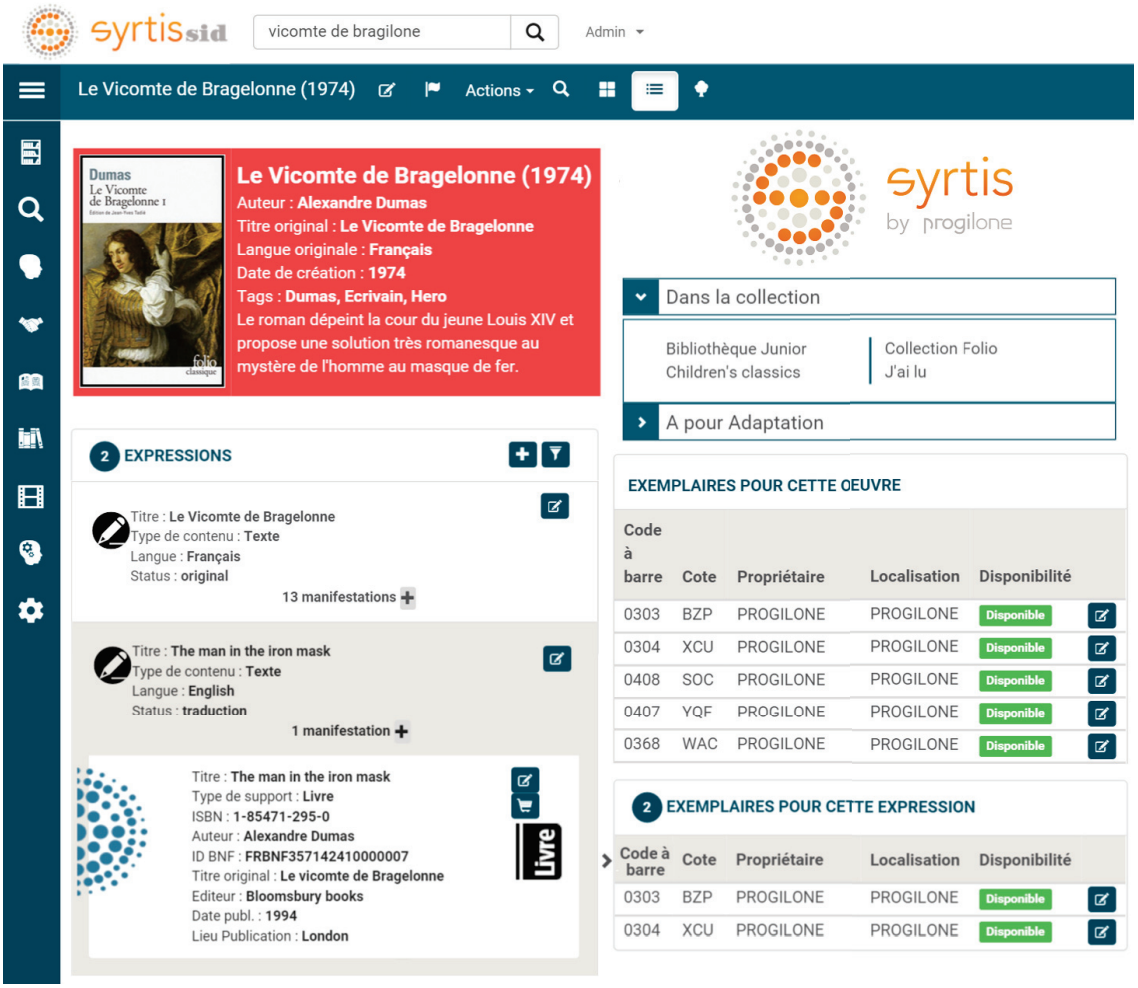


FIGURE 6.9 – Capture d'écran de la représentation d'une Œuvre FRBR dans Syrtis

Dans le cadre du projet Syrtis, COM3ET a été pré-configuré pour contenir les informations sur 460 champs issus de formats de type MARC (ici, MARC21 et UNIMARC). De plus, le méta-modèle des connaissances bibliographiques intègre 70 motifs de connaissances dont les instances, sous la forme d'un modèle de migration, intègrent 370 correspondances de migration (tous les champs MARC n'étant pas nécessairement utilisés) 30 conditions sur les motifs, 70 fonctions de transformation de valeurs de champs et 10 fonctions permettant des traitements plus lourd sur les champs. Le modèle utilisé dans les motifs utilise 11 classes (issues de FRBR) et 160 propriétés issues essentiellement de la norme RDA.

5. `https://www.progilone.fr/`

Évaluation préliminaire avec T42

Une expérimentation préliminaire a consisté à évaluer l'outil COM3ET avec le benchmark BIB-R et notamment le jeu de données T42 qui permet d'isoler les points faibles des outils de FR-BRisation. L'objectif consiste ici à s'assurer que l'implémentation de notre approche de méta-modélisation n'obtient pas un niveau de qualité inférieur aux solutions existantes. Plusieurs tests de T42 sont sélectionnés pour évaluer les résultats de FRBRisation de COM3ET. Les métriques d'évaluation choisies mesurent à la fois la complétude de la migration (MD-x) et la détection des motifs de connaissances bibbliographiques (MEND, MRND et ESE). La figure 6.10 présente ces résultats et les intègre avec les résultats des solutions évaluées dans le chapitre 4.

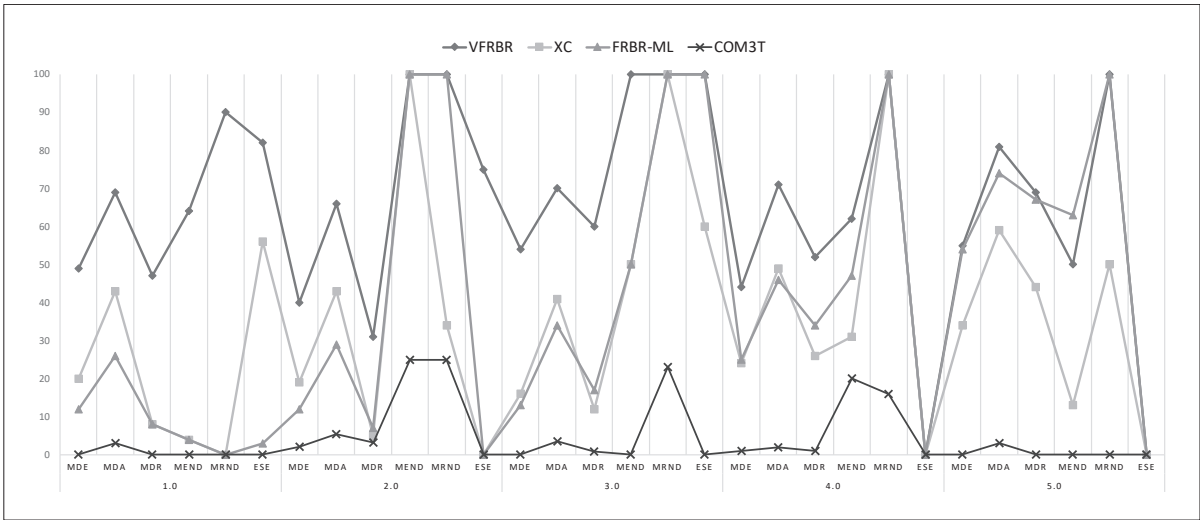


FIGURE 6.10 – Comparaison des données manquantes et erreurs d'interprétations des motifs de connaissances pour de la migration de tests de T42 avec les quatre outils évalués. Les résultats de COM3ET sont symbolisés par des croix.

Sur la figure 6.10, nous ne comparons les outils que sur des tests ne comportant pas d'anomalies dans T42 car nous nous concentrons sur l'interprétation des connaissances bibliographiques. Les résultats de COM3ET sont représentés par la courbe avec des croix. Nous observons que pour l'ensemble des tests, COM3ET ne dépasse pas les 30% d'erreurs. La solution COM3ET bénéficie d'une analyse des champs des notices, permettant d'activer automatiquement des correspondances dans le modèle de migration. Cette analyse permet de réduire les problèmes de complétude dans la migration des données.

Dans les expérimentations avec T42, les manquements en termes de complétude de COM3ET sont inférieurs à 5%. Concernant les erreurs principales de COM3ET sur les motifs, les augmentations sont bien détectées (car 0% d'erreurs en ESE), mais les entités et relations principales sont à 25% en erreur. Cela s'explique car certains choix de modélisation de l'augmentation dans T42 diffèrent de ceux de COM3ET. Par exemple, les illustrations sont parfois représentées comme des Œuvres à part entière dans T42 quand elles sont symbolisées par de nouvelles Expressions d'une Œuvre dans COM3ET. Les divergences de modélisation, constatées également sur l'outil XC lors de précédentes expérimentations, s'observent grâce aux métriques IAD et SMD. La figure 6.11 présente les résultats de COM3ET sur T42 selon ces deux métriques. Les taux d'IAD, lorsqu'ils dépassent 10% correspondent essentiellement à des propriétés ou des relations supplémentaires qui sont ajoutées dans les motifs de COM3ET mais qui ne sont pas explicités dans T42.

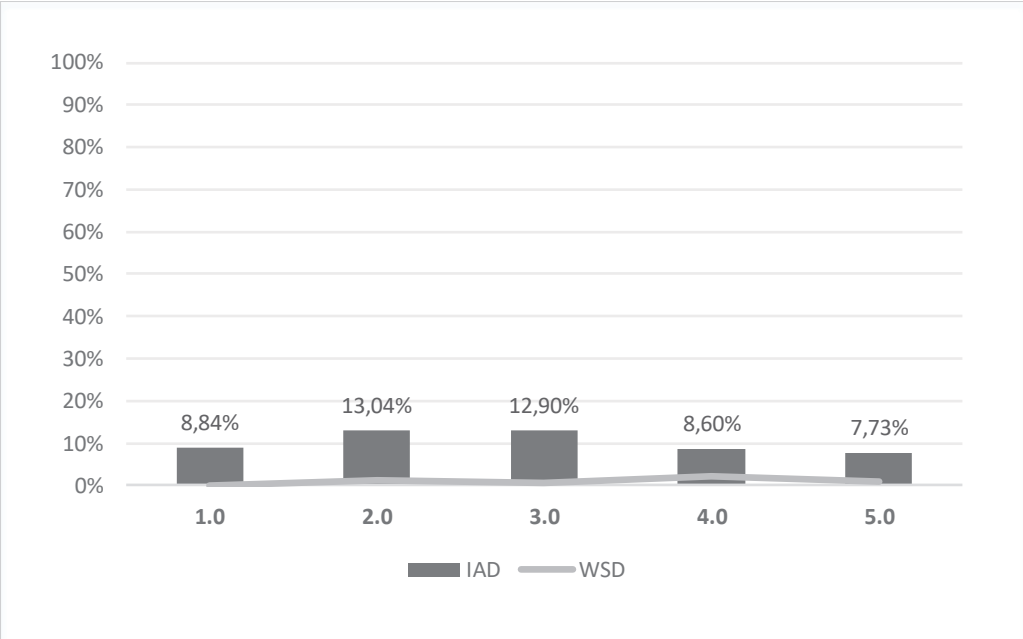


FIGURE 6.11 – Résultats de COM3ET sur T42 selon les métriques évaluant les données incorrectement ajoutées ou mal placées par rapport à l'expertise de T42.

Les résultats de cette expérimentation montrent que notre méthode de méta-modélisation obtient des scores intéressants en termes de qualité de la migration. Cela appuie les observations réalisées au chapitre 4, lors de l'expérimentation avec "FRBR-ML corrigé", que la prise en compte des spécificités du domaine bibliographique améliore la qualité de la migration. Par ailleurs, la possibilité de réutiliser des motifs de connaissances, dans COM3ET, qui ont déjà été documentés dans de précédents projets, permet de réduire sensiblement les efforts liés au traitement de nouveaux catalogues à migrer.

6.6.1 Projets réels de migration

Nous présentons maintenant plusieurs retours d'expériences sur des projets de migration de notices bibliographiques dans des contextes réels et industriels. L'objectif de cette partie consiste à mieux distinguer les problèmes, que nous avons évoqué sur ce processus, qui deviennent cruciaux lorsque l'on considère les contraintes réelles de ces projets. Nous décrivons dans la suite les projets de migrations réalisés avec le projet Syrtis.

Notices de lecture publique

Un premier catalogue sur lequel nous avons travaillé, que nous nommons *BMVV*, comporte environ 100.000 notices bibliographiques issues du domaine de la lecture publique. Il s'agit donc essentiellement de ressources documentaires courantes et non spécialisées (*ex.*, Harry Potter, Da Vinci Code). Ce type de bibliothèque ou médiathèque a une forte dépendance à la bibliothèque nationale, aussi, l'essentiel du travail de FRBRisation consiste ici à aligner les notices à traiter avec les données qui sont déjà "FRBRisées" dans le catalogue national. Dans le contexte de cette thèse, le catalogue FRBR de la bibliothèque nationale, *data.bnf*, n'en est qu'à ses débuts et est encore au stade d'expérimentation. C'est pourquoi un travail de FRBRisation, avec l'application d'un modèle de règles, est tout de même nécessaire. Toutefois, dans le contexte de la thèse, c'est une version simplifiée de COM3ET qui a été utilisée car un traitement sur les notices avait été déjà réalisé avec d'autres outils.

Une fois que le processus de migration a été réalisé, le catalogue FRBRisé obtenu comporte 70.000 œuvres. De nombreuses Œuvres représentent des créations isolées n'appartenant pas à des familles bibliographiques importantes, ou l'information sur la famille n'est pas représentée dans le catalogue initial. En conséquence le nombre d'Expressions par Œuvre est pour la grande majorité de 1. Nous notons cependant que 7400 Œuvres contiennent au moins un lien vers d'autres Œuvres. La nature de ces liens concerne essentiellement des relations agrégation ou d'adaptation. Nous avons cherché à valoriser ces relations avec de nouvelles interfaces comme une représentation en graphe de ces liens, illustrée sur la figure 6.12.

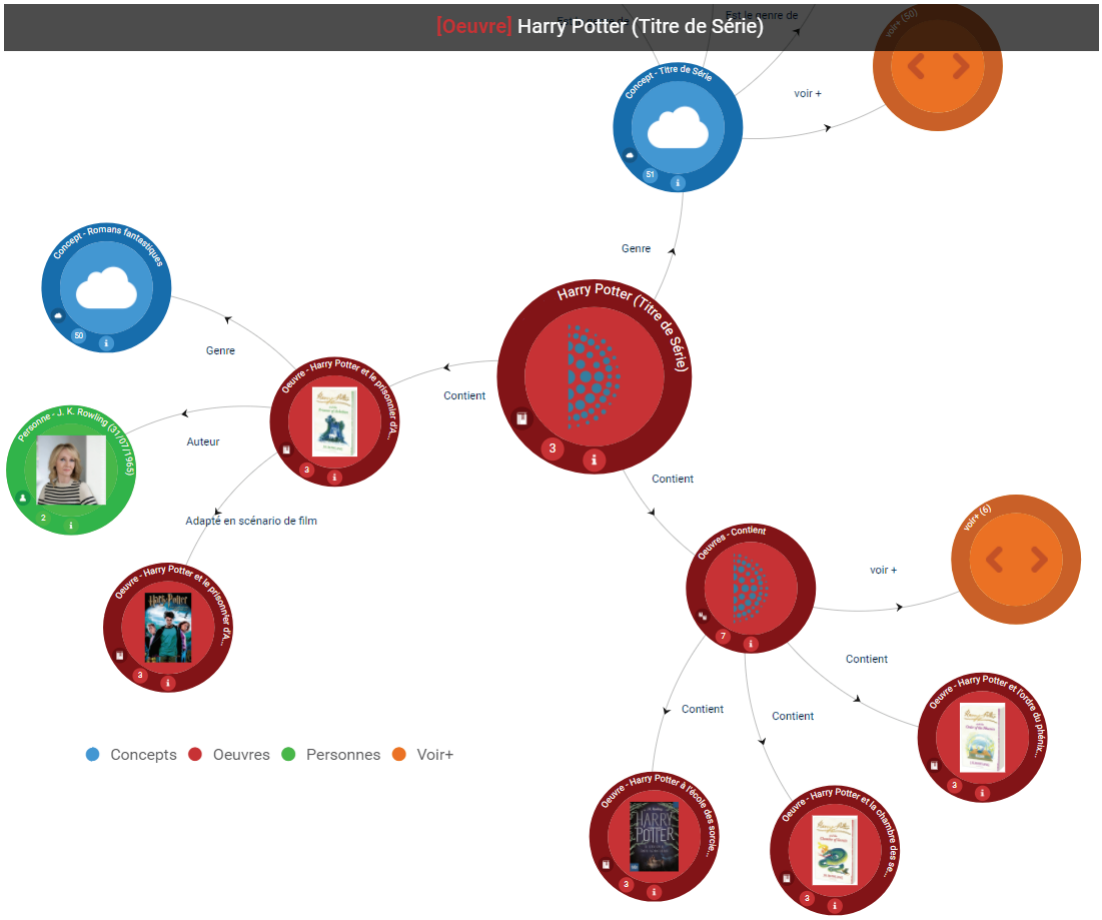


FIGURE 6.12 – Représentation en graphe d'entités FRBR dans Syrtis

Notre sentiment sur ce projet est qu'avec l'avancée des travaux des bibliothèques nationales sur la FRBRisation et le développement de nouvelles APIs pour la synchronisation des données entre institutions (*ex.*, SRU BNF [6]), la migration des notices de lecture publique demeure un processus simple qui peut être réalisé avec des outils existants.

Ressources documentaires spécialisées

Nous avons également travaillé à la migration de catalogues bibliographiques de ressources plus spécialisés. Le catalogue nommé HDL contient 400.000 notices décrivant des documents scientifiques sur le domaine de la médecine (articles, journaux, textes officiels). Le catalogue nommé CNP contient 200.000 notices de ressources du domaine pédagogiques (manuels scolaires, documents numériques). Ces catalogues contiennent essentiellement des œuvres isolées et qui ne sont

6. `http://www.bnf.fr/fr/professionnels/proto_sru/s.proto_sru_intro.html`

pas concernées par des motifs d'augmentation, de dérivation ou d'adaptation. Par ailleurs, seules 4000 Œuvres migrées ont plus de 1 Expression dans le catalogue HDL. Cependant, ces catalogues présentent deux particularités qui nécessitent des efforts supplémentaires qui sont la modélisation des connaissances spécifiques et l'intégration de multiples sources d'autorités externes. La modélisation spécifique des connaissances consiste en une adaptation des modèles conceptuels bibliographiques comme FRBR pour répondre à un fonctionnement spécifique d'une institution. Dans le cas d'HDL, les responsabilités des documents sont majoritairement des scientifiques qui publient des travaux dans le cadre d'une affiliation avec un établissement de recherche. Cette affiliation fait intervenir une nouvelle entité dans le modèle entre une Œuvre et un Agent. Dans le cas du catalogue CNP, des agrégations spécifiques entre les œuvres viennent répondre à une organisation intellectuelle spécifique.

Les autorités sont les entités (*ex.*, Concepts, Agents) qui décrivent le contexte intellectuel des Œuvres d'un catalogue. Dans la pratique courante, les noms associés à ces entités sont conservés et gérés dans des listes de valeurs annexes au catalogue bibliographique. Dans le cas de ressources plus spécialisés, les Œuvres sont décrites par de nombreuses listes de valeurs qui peuvent contenir des hiérarchies complexes pour organiser les termes à décrire (*ex.*, thésaurus des maladies pour HDL, thésaurus des disciplines scolaires pour CNP). De plus, dans le contexte d'une adoption des technologies du Web Sémantique, ces vocabulaires doivent aussi migrer vers des formats (*ex.*, RDF) et modèles (*ex.*, SKOS) facilitant l'interopérabilité avec d'autres catalogues. C'est pourquoi la migration de tels catalogues s'accompagne également d'un travail important de normalisation des listes de valeurs et thésaurus en réalisant notamment des alignements via des outils d'Ontology Matching. Dans le contexte de nos travaux, nous avons par exemple utilisé l'outil *AgreementMaker*, [46], pour aligner 6 thésaurus médicaux afin de faciliter les enrichissement et la navigation entre les ressources du catalogue HDL.

Migration d'un grand volume de notices

Dans un processus complet de migration de notices bibliographiques, les phases d'analyse des données, de nettoyage de ces dernières, de déduplication des entités ou d'alignement de ces dernières avec des sources externes augmente le temps global de traitement. Nous avons vu notamment au chapitre 6 des solutions pour améliorer les performances de la déduplication grâce à la phase de blocking dont nous avons également montré comment l'automatiser. Notre outil COM3ET intègre ces solutions ainsi que d'autres améliorations comme la structure en arbre des motifs de connaissances avec une hiérarchie des conditions permettant d'éviter des calculs inutiles, dans les motifs, si la condition globale du motif n'est pas validée. Nous avons expérimenté ces améliorations sur un catalogue, nommé DEC, d'environ 2.000.000 de notices. La migration de DEC avec un jeu de règles standard dans COM3ET produit 11.000.000 millions d'entités FRBR dont 1,7 millions d'œuvres. Un premier temps de traitement a été relevé suite à une application de COM3ET sur le catalogue sans ajustement particuliers. Ce temps initiale était de 20h. Nous avons réalisé des améliorations dans le processus afin de réduire ce temps à 8h. Nous décrivons ces améliorations ci-après.

Ces expérimentations nous ont permis d'observer notamment certaines limitations de notre implémentation de déduplication. En effet, la méthode de blocking basée sur *PairRange* (présentée au chapitre 3) repose sur deux phases consistant successivement à isoler les attributs permettant de générer des blocs d'entités à comparer puis à énumérer les pairs dans ces blocs afin d'homogénéiser la taille des blocs. Malheureusement, une grande partie des ressources décrites sont isolées et peu d'informations permettent de générer des clefs de blocking efficaces. Ainsi, si les blocs homogènes permettent effectivement de réaliser un calcul parallèle efficace des comparaisons d'entités, la phase d'initialisation de la matrice d'énumération demande un temps de

traitement considérable. Il est dans ce cas nécessaire d'apporter préalablement d'autres critères qui permettent de discréditer en amont certaines paires. Par exemple, nous avons comparé initialement l'égalité (faible en temps de calcul) des chaînes de caractères composant le nom des responsabilités des œuvres afin de décider de ne pas comparer les entités dont aucune chaîne ne convenait (en omettant les particules dans le nom).

L'application des règles a également nécessité de créer des motifs "abstraits", c'est à dire de nouveaux nœuds "vides" dans l'arbre du modèle de migration, permettant d'augmenter la granularité des conditions appliquées aux notices et réduire encore le nombre de traitements inutiles. Par exemple, un motif a été créé en amont des différents motifs liés au sujet d'une Œuvre FRBR pour vérifier si les données sur le sujet étaient conformes et pouvaient être récupérées. Enfin, nous avons dû ajuster les outils qui stockent les données migrées pendant le processus afin d'éviter des accès disques trop importants. Nous avons par exemple utilisé le moteur d'indexation d'Elasticsearch [7] pour stocker temporairement des entités migrées dans l'objectif de les réutiliser afin de créer des liens avec d'autres entités issues de notices traitées ultérieurement dans le projet.

6.7 Conclusion

Dans ce chapitre, nous avons présenté **CoM3ET** qui est une implémentation technique d'un système d'intégration de métadonnées bibliographiques. Plus particulièrement nous avons décrit deux phases essentielles du système qui sont la création du modèle de migration et d'enrichissement et la déduplication des entités. Ce travail nous permet de montrer les possibilités d'automatisation de ces phases, tout en respectant les exigences de qualité du domaine documentaire. Nous avons montré qu'il est possible de rendre la migration des notices complètement automatique, à condition que les spécificités des notices à traiter soient préalablement configurées. Dans ce sens, nous préconisons une description claire des motifs de connaissances afin de pouvoir les réutiliser dans différents projets.

7. https ://www.elastic.co/fr/

Chapitre 7

Conclusion

Ce chapitre conclut les travaux réalisés dans cette thèse. Les recherches présentées dans ce manuscrit sont focalisées sur l'intégration de métadonnées bibliographiques et la qualité des bases de connaissances documentaires. Nos contributions concernent les tâches de modélisation, de migration et d'enrichissement des entités bibliographiques.

Nous avons considéré le double enjeu consistant d'une part à faciliter la migration et l'enrichissement des notices bibliographiques pour les institutions documentaires et d'autre part à considérer les exigences de qualité du domaine dans un contexte d'évolution des normes. Nous avons également étudié les solutions existantes dans ce contexte et les attentes persistantes des acteurs du domaine pour mieux cibler les efforts nécessaires à une adoption plus importante des technologies du Web Sémantique dans les institutions documentaires. Notre hypothèse principale est que les spécificités des données bibliographiques, comme les relations intellectuelles et éditoriales entres les Œuvres, doivent être prises en compte lors de la création et la configuration des outils d'intégration de métadonnées. La contribution principale de cette thèse consiste ainsi en une méthodologie de modélisation des règles de migration et d'enrichissement des métadonnées en tenant compte des métriques de qualité du domaine. Le projet Syrtis a offert un cadre industriel à cette thèse, nous permettant d'évaluer notre approche dans un contexte concret et d'ainsi relever des problématiques récurrentes du domaine. Les choix d'évaluation et d'implémentation on donc été influencé, à la fois par le besoin de fixer et mieux appréhender les critères et méthodes dans la communauté, et de répondre à des problématiques au cœur des discussions et travaux dans le domaine documentaire.

Dans le reste de ce chapitre, nous rappelons les contributions principales de la thèse, puis nous discutons des travaux futurs qui restent à entreprendre. Enfin, nous présentons certaines perspectives pour l'évolution des systèmes d'information bibliographiques.

7.1 Contributions

Cette thèse commence par présenter un état des lieux de la transition bibliographique dans la communauté documentaire. Les chapitre 2 et 3 livrent successivement les particularités du domaine documentaire ainsi que les solutions proposées pour transformer d'anciens catalogues en de nouvelles bases de connaissances sémantiques. En particulier, nous rassemblons sur un même plan les caractéristiques des connaissances bibliographiques, notamment des relations entre les œuvres formant des familles bibliographiques, et les problématiques de modélisation des solutions de migration des notices. Ce travail nous permet de mettre en lumière les difficultés d'interprétation que peuvent impliquer les données bibliographiques qui décrivent, de manière implicite, des relations riches entre les documents. Nous continuons par proposer une classification des

solutions de FRBRisation afin d'identifier les forces et faiblesses des outils existants. Cette étude nous permet de mieux appréhender les problèmes techniques du domaine en observant les fonctionnalités spécifiques de ces outils. De plus, cela nous permet de comprendre l'évolution des techniques de FRBRisation dans un contexte d'apprentissage, par la communauté, des nouvelles normes documentaires qui intègrent les principes du Web Sémantique.

Nous orientons ensuite nos travaux sur les problématiques de qualité posées par la transformation des notices existantes. Nous supposons d'ailleurs que ces problématiques sont la cause du manque d'adoption massives des modèles et principes du web de données dans l'univers bibliographique. Nous étudions, dans le chapitre 4, les critères de qualité des bases de connaissances bibliographiques et nous proposons de nouvelles métriques qui tiennent compte des spécificités du domaine. Ces métriques permettent à la fois d'évaluer les règles dédiées à l'interprétation des notices documentaires, le processus de migration lui-même et enfin les résultats de ce dernier. Nous complétons plus tard ces travaux en proposant de nouveaux jeux de données permettant d'évaluer la transformation des notices bibliographiques. Ces jeux de données sont systématiquement composés d'une collection à migrer ainsi que d'une proposition d'expertise réalisée manuellement et validée par des experts. L'ensemble des métriques et jeux de données forment le benchmark que nous appelons *BIB-R* et qui est le premier benchmark spécialisé sur l'interprétation et la transformation des notices bibliographiques.

Ayant constaté des attentes qualitatives du domaine et ayant évalué certains outils existants grâce au benchmark BIB-R, nous proposons une nouvelle méthodologie permettant la création d'un modèle de migration pouvant satisfaire les attentes des institutions documentaires. Le chapitre 5 décrit cette méthodologie qui s'applique successivement au problème de la transformation des métadonnées bibliographiques en entités et relations puis à l'enrichissement de ces dernières avec des sources externes. Nos propositions se basent notamment sur une représentation des règles de migration et d'enrichissement à un plus haut niveau d'abstraction que l'existant afin de faciliter la prise en main des règles par les experts du domaine. De plus, notre organisation des règles en graphe de motifs de connaissances offre différents bénéfices pour l'application automatique des règles, en évitant des calculs inutiles selon la hiérarchie des motifs, et la maintenance des ces règles grâce à une limitation de la redondance dans le modèle. Nous présentons systématiquement une version formalisée de notre méthodologie avec des exemples d'application.

Nous poursuivons, au chapitre 6, par une proposition d'implémentation pratique de notre approche théorique du modèle de migration. Nous décrivons ainsi l'outil COM3ET qui permet la migration de grands volumes de notices vers une base de connaissances sémantiques. Cette dernière, une fois créée, peut être connectée à des sources externes pour réaliser un enrichissement des connaissances *a posteriori*. Notre implémentation intègre nativement une représentation en un arbre des motifs de connaissances bibliographiques, eux même couplés aux métriques de qualité présentées au chapitre 4. Cela nous permet d'obtenir de meilleurs résultats que l'existant, lors de la migration de catalogues bibliographiques, en termes de qualité du résultat et d'efforts pour maintenir ou adapter les règles.

7.2 Travaux futurs

Les avancées en matière de migration et d'enrichissement de données facilitent le mouvement vers l'adoption des technologies du Web Sémantique dans les institutions culturelles [147]. Cependant, cette avancée ouvre de nouveaux enjeux comme la pérennisation des entités nouvellement créées. En effet, les institutions ayant un rôle d'éditeur de métadonnées bibliographiques doivent désormais créer des identifiants pérennes, notamment concernant les œuvres bibliographiques FRBR,

qui n'avaient pas d'équivalent dans les anciens catalogues. Ces identifiants comme l'ARK [1], qui tend à devenir un standard dans le monde documentaire, facilite la création de liens RDF entre les entités et évite la redondance dans les bases de connaissances. Aussi, nous prévoyons d'intégrer cette construction d'identifiants dans nos outils de migration de catalogues.

La migration des notices vers des bases de connaissances sémantiques devient cruciale dans la communauté documentaire et les projets dédiés à ce processus vont probablement se multiplier dans les années à venir. Lorsqu'une majorité de catalogues seront transformés, l'essentiel des tâches futurs va concerner l'alignement et la déduplication des entités bibliographiques avec d'autres bases de connaissances. Dans ce contexte, les processus d'alignements de schémas (Ontology Matching) et d'entités (Entity Matching) seront probablement (re)mis au centre des réflexions dans la communauté. Toutefois, les équivalences qui sont progressivement détectées au niveau des entités doivent être réutilisées pour améliorer les processus d'alignement de schéma [36]. Dans le contexte des solutions basées sur la technique d'*Instance-based Ontology Matching*, la prise en compte des nouveaux alignements entre les entités bibliographiques peut améliorer les performances et la qualité de ces solutions [141]. Dans nos travaux nous prévoyons d'évaluer cet impact en réalisant des expérimentations pour l'alignement de modèles basés sur FRBR/LRM.

7.3 Perspectives à plus long terme

Une première perspective pour nos travaux réside dans une amélioration du processus d'enrichissement des entités bibliographiques par la détection automatique de nouveaux motifs de connaissances dans les bases de données où les informations bibliographiques sont implicitement représentées. Plus particulièrement, nous cherchons à détecter des correspondances one-to-many entre les sources, c'est à dire le cas où une propriété sémantique dans une source A est équivalente à une combinaison de propriétés dans une source B. La figure 7.1 présente un exemple illustrant cette problématique.

Concrètement, comme illustré par la Figure 7.1, la propriété *estUneAdaptation* dans la source A permet de mettre en relation l'entité E_1 symbolisant l'Œuvre "Germinal" d'Emile Zola (un Agent représenté par l'entité E_0) et l'entité E_2, représentant le film "Germinal" réalisé par Claude Berry en 1993 (toujours dans la source A). En considérant une source externe B, nous pourrions enrichir la connaissance liée à l'entité E_1 en la reliant à l'entité E'_3 de la source B représentant le film "Germinal" réalisé par Yves Allégret en 1962. Cependant, la propriété *estUneAdaptationDe* n'est pas définie dans le schéma de la source B. Toutefois, dans la source B, il existe l'entité E'_1 représentant l'Œuvre "Germinal" d'Emile Zola qui est en lien avec l'entité E'_2 représentant le réalisateur Yves Allégret par la propriété *aFaitUneAdaptation*. L'entité E'_2 est en lien avec la propriété E'_3 par la propriété *aRealise*. Pour pouvoir enrichir l'entité E_1, nous procédons à une étape d'*Entity Linking* qui permet de mettre en relation l'entité E_1 avec l'entité E'_1. Ensuite, il est nécessaire de trouver que la propriété *estUneAdaptationDe* de la source A peut être déduite par la composition des propriétés *aFaitUneAdaptation* et *aRéalisé*.

Une autre perspective concerne l'amélioration de la pertinence des moteurs de recherche quand les critères de recherche de l'utilisateur sont abstraits ou flous (*ex.*, recherche sur le "corps humain"). L'opportunité ici est que l'augmentation des entités bibliographiques, des agents associés et des concepts sujets (des Œuvres) dans le web de données, grâce à la FRBRisation des notices, offre de nouvelles possibilités pour combiner ces concepts voir en associer de nouveaux aux entités bibliographiques. L'idée est donc de profiter de cette émergence de concepts sujets qui sont associés aux Œuvres bibliographiques pour améliorer l'expérience utilisateur dans la recherche.

1. `http://www.bnf.fr/fr/professionnels/issn_isbn_autres_numeros/a.ark.html`

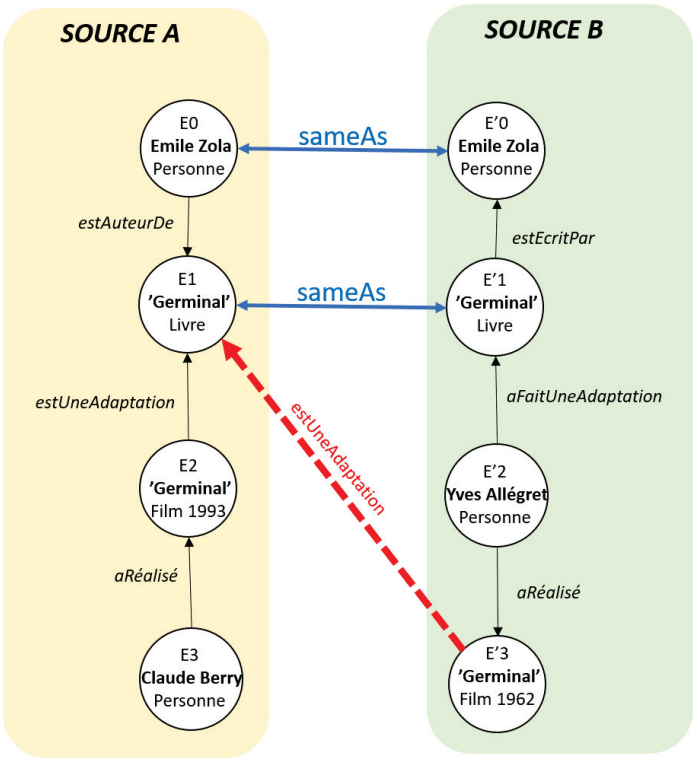


FIGURE 7.1 – Exemple de découverte d'une connaissances bibliographique implicite

La figure 7.2 présente plusieurs processus qui peuvent être intégrés à un moteur de recherche pour l'améliorer.

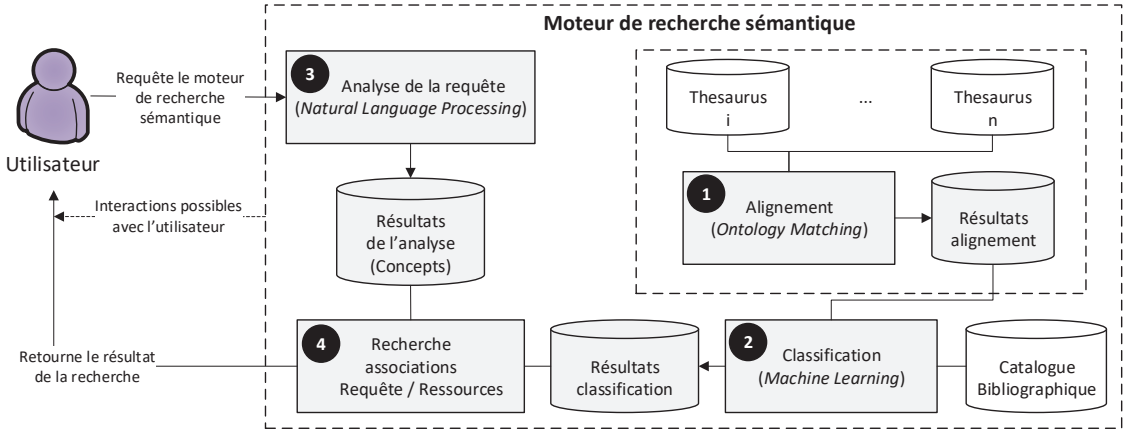


FIGURE 7.2 – Phases internes pour un moteur de recherche sémantique

Dans ce prototype de moteur de recherche, l'idée est d'associer des thésaurii de concepts avec des entités bibliographiques (catalogue bibliographique sur la figure 7.2) pour enrichir les Œuvres bibliographiques de nouveaux termes issus de l'alignement des thésaurii. Il est également possible de créer des méta-concepts issus d'un processus de classification qui pourrait extraire ces méta-concepts grâce à l'évaluation de la fréquence dans laquelle certains concepts sujets se retrouvent connectés aux mêmes œuvres bibliographiques. Cela revient à un algorithme de *clustering*[2], basé

2. `https://fr.wikipedia.org/wiki/Partitionnement_de_données`

sur la fréquence (mais d'autres critères pourraient être pertinent) des concepts sujets d'Œuvres bibliographiques dans un contexte de multiples sources documentaires et de multiples thésaurii. L'objectif final est d'obtenir un foisonnement de concepts qui permettent de mieux interpréter la requête des utilisateurs et de les impliquer dans le processus de recherche en leur proposant de spécifier la sémantique des termes qu'ils ont saisis dans le moteur de recherche [63].

Bibliographie

[1] Trond Aalberg. A Process and Tool for the Conversion of MARC Records to a Normalized FRBR Implementation. *International Conference on Asian Digital Libraries*, 2006.

[2] Trond Aalberg, Tanja Merčun, and Maja Žumer. Coding FRBR-structured bibliographic information in MARC. In *Digital Libraries : For Cultural Heritage, Knowledge Dissemination, and Future Creation*, pages 128–137. Springer, 2011.

[3] Trond Aalberg, Tanja Merčun, and Maja Žumer. Interactive displays for the next generation of entity-centric bibliographic models. In *International Conference on Asian Digital Libraries*, pages 199–211. Springer, 2017.

[4] Trond Aalberg, Audun Vennesland, and Maliheh Farrokhnia. A Pattern-Based Framework for Best Practice Implementation of CRM/FRBRoo. In *East European Conference on Advances in Databases and Information Systems*, pages 438–447. Springer, 2015.

[5] Trond Aalberg and Maja Žumer. The value of MARC data, or, challenges of frbrisation. *Journal of Documentation*, 69(6) :851–872, 2013.

[6] Manel Achichi, Rodolphe Bailly, Cécile Cecconi, Marie Destandau, Konstantin Todorov, and Raphaël Troncy. Doremus : Doing reusable musical data. In *ISWC : International Semantic Web Conference*, 2015.

[7] Maristella Agosti, Nicola Ferro, and Gianmaria Silvello. Digital library interoperability at high level of abstraction. *Future Generation Computer Systems*, 55 :129–146, 2016.

[8] Getaneh Alemu, Brett Stevens, and Penny Ross. Towards a conceptual framework for user-driven semantic metadata interoperability in digital libraries : A social constructivist approach. *New Library World*, 113(1/2) :38–54, 2012.

[9] Getaneh Alemu, Brett Stevens, Penny Ross, and Jane Chandler. Linked Data for libraries : Benefits of a conceptual shift from library-specific record structures to RDF-based data models. *New Library World*, 113(11/12) :549–570, 2012.

[10] Anila Angjeli, Bertrand Caron, and Emmanuelle Bermès. Data processing for digital libraries : the experience of the BnF with Europeana Sounds project. In *IFLA World Library and Information Conference, 82nd IFLA General Conference and Assembly, satellite pre-conference" Data in libraries : the big picture"*, 2016.

[11] Marie-Louise Ayres. Case studies in implementing functional requirements for bibliographic records [FRBR] : AustLit and MusicAustralia. *The Australian library journal*, 54(1) :43–54, 2005.

[12] Franz Baader. *The description logic handbook : Theory, implementation and applications*. Cambridge university press, 2003.

[13] Alison Babeu. Building a "FRBR-inspired" catalog : The Perseus digital library experience. *Retrieved November*, 17 :2008, 2008.

[14] Thomas Baker. Libraries, languages of description, and linked data : a Dublin Core perspective. *Library Hi Tech*, 30(1) :116–133, 2012.

[15] Zohra Bellahsene, Angela Bonifati, Fabien Duchateau, and Yannis Velegrakis. On evaluating schema matching and mapping. In *Schema matching and mapping*, pages 253–291. Springer, 2011.

[16] Rick Bennett, Brian F Lavoie, and Edward T O'neill. The concept of a work in WorldCat : an application of FRBR. *Library Collections, Acquisitions, and Technical Services*, 27(1) :45–59, 2003.

[17] Philip A Bernstein and Sergey Melnik. Model management 2.0 : manipulating richer mappings. In *Proceedings of the 2007 ACM SIGMOD international conference on Management of data*, pages 1–12. ACM, 2007.

[18] Mikhail Bilenko, Beena Kamath, and Raymond J Mooney. Adaptive blocking : Learning to scale up record linkage. In *Data Mining, 2006. ICDM'06. Sixth International Conference on*, pages 87–96. IEEE, 2006.

[19] Jens Bleiholder and Felix Naumann. Data fusion. *ACM Computing Surveys (CSUR)*, 41(1) :1, 2009.

[20] Jennifer B Bowen. Moving Library Metadata toward Linked Data : Opportunities Provided by the eXtensible Catalog. In *International Conference on Dublin Core and Metadata Applications*, pages 44–59, 2010.

[21] Scott Britell, Lois M L Delcambre, and Paolo Atzeni. Facilitating Data-Metadata Transformation by Domain Specialists in a Web-Based Information System Using Simple Correspondences. In *Conceptual Modeling : 35th International Conference, ER 2016, Gifu, Japan, November 14-17, 2016, Proceedings 35*, pages 445–459. Springer, 2016.

[22] D Grant Campbell. The Dire Straits of Bibliographic Families. In *Proceedings of the Annual Conference of CAIS/Actes du congrès annuel de l'ACSI*, 2013.

[23] Gustavo Candelaa, Pilar Escobarb, Rafael C Carrascob, and Manuel Marco-Suchb. Migration of a library catalogue into RDA linked open data. *Semantic Web journal*, 2015.

[24] Allyson Carlyle. User categorisation of works : Toward improved organisation of online catalogue displays. *Journal of documentation*, 55(2) :184–208, 1999.

[25] Naicheng Chang, Yuchin Tsai, Gordon Dunsire, and Alan Hopkinson. Experimenting with implementing FRBR in a Chinese Koha system. *Library Hi Tech News*, 30(10) :10–20, 2013.

[26] Angelo Chianese, Fiammetta Marulli, Francesco Piccialli, and Isabella Valente. A novel challenge into multimedia cultural heritage : An integrated approach to support cultural information enrichment. In *Proceedings - 2013 International Conference on Signal-Image Technology and Internet-Based Systems, SITIS 2013*, pages 217–224. IEEE, 2013.

[27] Peter Christen. A survey of indexing techniques for scalable record linkage and deduplication. *IEEE transactions on knowledge and data engineering*, 24(9) :1537–1555, 2012.

[28] Peter Christen. *Data matching : concepts and techniques for record linkage, entity resolution, and duplicate detection.* Springer Science & Business Media, 2012.

[29] William Cohen, Pradeep Ravikumar, and Stephen Fienberg. A comparison of string metrics for matching names and records. In *Kdd workshop on data cleaning and object consolidation*, volume 3, pages 73–78, 2003.

[30] Karen Coyle. The library catalog in a 2.0 world. *The Journal of Academic Librarianship*, 33(2) :289–291, 2007.

[31] Karen Coyle. FRBR, twenty years on. *Cataloging & Classification Quarterly*, 53(3-4) :265–285, 2015.

[32] Karen Coyle and Diane Hillmann. Resource description and access (RDA). *D-Lib magazine*, 13(1/2) :1082–9873, 2007.

[33] Gianfranco Crupi. Beyond the Pillars of Hercules : Linked data and cultural heritage. *JLIS.it*, 4(1), 2013.

[34] Hong-Jie Dai, Chi-Yang Wu, R Tsai, Wen-Lian Hsu, and Others. From entity recognition to entity linking : a survey of advanced entity linking techniques. In *The 26th Annual Conference of the Japanese Society for Artificial Intelligence*, pages 1–10, 2012.

[35] Joffrey Decourselle, Fabien Duchateau, and Nicolas Lumineau. A Survey of FRBRization Techniques. In *Theory and Practice of Digital Libraries (TPDL)*, pages 185–196, 2015.

[36] Joffrey Decourselle, Audun Vennesland, Trond Aalberg, Fabien Duchateau, and Nicolas Lumineau. A novel vision for navigation and enrichment in cultural heritage collections. In *East European Conference on Advances in Databases and Information Systems*, pages 488–497. Springer, 2015.

[37] Timothy J Dickey. FRBRization of a library catalog : better collocation of records, leading to enhanced search, retrieval, and display. *Information Technology and Libraries*, 27(1) :23, 2008.

[38] Xin Luna Dong, Evgeniy Gabrilovich, Geremy Heitz, Wilko Horn, Kevin Murphy, Shaohua Sun, and Wei Zhang. From data fusion to knowledge fusion. *Proceedings of the VLDB Endowment*, 7(10) :881–892, 2014.

[39] Uwe Draisbach and Felix Naumann. A comparison and generalization of blocking and windowing algorithms for duplicate detection. In *Proceedings of the International Workshop on Quality in Databases (QDB)*, pages 51–56, 2009.

[40] René-Vincent du Grandlaunay. L'application AlKindi - FRBR-FRAD et RDA - Au service de la rencontre interculturelle et interreligieuse. In *Libraries Serving Dialogue*, pages 91–111. 2014.

[41] Arnab Dutta, Christian Meilicke, and Heiner Stuckenschmidt. Enriching Structured Knowledge with Open Information. In *Proceedings of the 24th International Conference on World Wide Web - WWW '15*, pages 267–277. International World Wide Web Conferences Steering Committee, 2015.

[42] Ahmed K Elmagarmid, Panagiotis G Ipeirotis, and Vassilios S Verykios. Duplicate record detection : A survey. *IEEE Transactions on knowledge and data engineering*, 19(1) :1–16, 2007.

[43] John L Espley and Robert Pillow. The VTLS implementation of FRBR. *Cataloging & classification quarterly*, 50(5-7) :369–386, 2012.

[44] Oren Etzioni, Michele Banko, Stephen Soderland, and Daniel S Weld. Open information extraction from the web. *Communications of the ACM*, 51(12) :68–74, 2008.

[45] Ronald Fagin, Laura M. Haas, Mauricio Hernández, Renée J. Miller, Lucian Popa, and Yannis Velegrakis. Clio : Schema mapping creation and data exchange. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 5600 LNCS :198–236, 2009.

[46] Daniel Faria, Catia Pesquita, Emanuel Santos, Matteo Palmonari, Isabel F Cruz, and Francisco M Couto. The agreementmakerlight ontology matching system. In *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"*, pages 527–541. Springer, 2013.

[47] Javier D Fernández, Wouter Beek, Miguel A Mart\'\inez-Prieto, and Mario Arias. LOD-a-lot : A queryable dump of the LOD cloud. 2017.

[48] Javier D Fernández, Miguel A Mart\'\inez-Prieto, Claudio Gutiérrez, Axel Polleres, and Mario Arias. Binary RDF representation for publication and exchange (HDT). *Web Semantics : Science, Services and Agents on the World Wide Web*, 19 :22–41, 2013.

[49] Nuno Freire, José Borbinha, and Pável Calado. Identification of FRBR works within bibliographic databases : An experiment with UNIMARC and duplicate detection techniques. *Asian Digital Libraries. Looking Back 10 Years and Forging New Frontiers*, pages 267–276, 2007.

[50] Francesca Frontini, Carmen Brando, and Jean-Gabriel Ganascia. Semantic web based named entity linking for digital humanities and heritage texts. In *First International Workshop Semantic Web for Scientific Heritage at the 12th ESWC 2015 Conference*, 2015.

[51] Norbert Fuhr, Giannis Tsakonas, Trond Aalberg, Maristella Agosti, Preben Hansen, Sarantos Kapidakis, Claus-Peter Klas, László Kovács, Monica Landoni, András Micsik, and Others. Evaluation of digital libraries. *International Journal on Digital Libraries*, 8(1) :21–38, 2007.

[52] Ted Gemberling. FRSAD, Semiotics, and FRBR-LRM. *Cataloging & Classification Quarterly*, 54(2) :136–144, 2016.

[53] Mauro Guerrini and Tiziana Possemato. Linked data : a new alphabet for the semantic web. *JLIS. it*, 4(1) :67, 2013.

[54] Mariá Hallo, Sergio Luján-Mora, Alejandro Maté, and Juan Trujillo. Current state of Linked Data in digital libraries. *Journal of Information Science*, page 0165551515594729, 2015.

[55] James A Hammerton, Michael Granitzer, Dan Harvey, Maya Hristakeva, and Kris Jack. On generating large-scale ground truth datasets for the deduplication of bibliographic records. In *Proceedings of the 2nd International Conference on Web Intelligence, Mining and Semantics*, page 18. ACM, 2012.

[56] Michael Hartung, Anika Groß, and Erhard Rahm. COnto–Diff : generation of complex evolution mappings for life science ontologies. *Journal of biomedical informatics*, 46(1) :15–32, 2013.

[57] Bernhard Haslhofer, Elaheh Momeni, Manuel Gay, and Rainer Simon. Augmenting Europeana content with linked data resources. In *Proceedings of the 6th International Conference on Semantic Systems - I-SEMANTICS '10*, page 1. ACM, 2010.

[58] Knut Hegna and Eeva Murtomaa. Data mining MARC to find : FRBR ? *International cataloguing and bibliographic control*, 32(3) :52–55, 2003.

[59] Mauricio A Hernández, Paolo Papotti, and Wang-Chiew Tan. Data exchange with data-metadata translations. *Proceedings of the VLDB Endowment*, 1(1) :260–273, 2008.

[60] Thomas B Hickey, Edward T O'Neill, and Jenny Toves. Experiments with the IFLA functional requirements for bibliographic records (FRBR). *D-Lib magazine*, 8(9) :1–13, 2002.

[61] Thomas Butler Hickey and Jenny Toves. *FRBR Work-Set Algorithm*. OCLC, 2005.

[62] Annika Hinze, David Bainbridge, Sally Jo Cunningham, and J Stephen Downie. Low-cost Semantic Enhancement to Digital Library Metadata and Indexing : Simple Yet Effective Strategies. In *Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries*, pages 93–102. ACM, 2016.

[63] Annika Hinze, Craig Taube-Schock, David Bainbridge, Rangi Matamua, and J Stephen Downie. Improving Access to Large-scale Digital Libraries ThroughSemantic-enhanced Search and Disambiguation. In *Proceedings of the 15th ACM/IEEE-CE on Joint Conference on Digital Libraries*, pages 147–156. ACM, 2015.

[64] Martin Höffernig, Werner Bailer, Günter Nagler, and Helmut Mülner. Mapping audiovisual metadata formats using formal semantics. In *International Conference on Semantic and Digital Media Technologies*, pages 80–94. Springer, 2010.

[65] Aidan Hogan, Antoine Zimmermann, Jürgen Umbrich, Axel Polleres, and Stefan Decker. Scalable and distributed methods for entity matching, consolidation and disambiguation over linked data corpora. *Web Semantics : Science, Services and Agents on the World Wide Web*, 10 :76–110, 2012.

[66] Ekaterini Ioannou, Nataliya Rassadko, and Yannis Velegrakis. On generating benchmark data for entity matching. *Journal on Data Semantics*, 2(1) :37–56, 2013.

[67] Yu Jiang, Can Lin, Weiyi Meng, Clement Yu, Aaron M Cohen, and Neil R Smalheiser. Rule-based deduplication of article records from bibliographic databases. *Database*, 2014, 2014.

[68] Ernesto Jiménez-Ruiz and Bernardo Cuenca Grau. Logmap : Logic-based and scalable ontology matching. In *International Semantic Web Conference*, pages 273–288. Springer, 2011.

[69] A Juffinger, E Lex, and N Freire. Error Tolerant Large Scale FRBRization. In *Proceeding of the Very Large Database Workshop 2011*, pages 47–59. ., 2011.

[70] F Tim Knight. Break on through to the Other Side : The Library & Linked Data. *TALL Q.*, 30 :6, 2011.

[71] Craig Knoblock, Pedro Szekely, José Ambite, Aman Goel, Shubham Gupta, Kristina Lerman, Maria Muslea, Mohsen Taheriyan, and Parag Mallick. Semi-automatically mapping structured sources into the semantic web. *The Semantic Web : Research and Applications*, pages 375–390, 2012.

[72] Craig A Knoblock, Pedro Szekely, Eleanor Fink, Duane Degler, David Newbury, Robert Sanderson, Kate Blanch, Sara Snyder, Nilay Chheda, Nimesh Jain, and Others. Lessons Learned in Building Linked Data for the American Art Collaborative. In *International Semantic Web Conference*, pages 263–279. Springer, 2017.

[73] Phokion G Kolaitis. Schema mappings, data exchange, and metadata management. In *Proceedings of the twenty-fourth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 61–75. ACM, 2005.

[74] Lars Kolb, Andreas Thor, and Erhard Rahm. Load balancing for mapreduce-based entity resolution. In *ICDE*, pages 618–629, 2012.

[75] Hanna Köpcke, Andreas Thor, and Erhard Rahm. Evaluation of entity resolution approaches on real-world match problems. *Proceedings of the VLDB Endowment*, 3(1-2) :484–493, 2010.

[76] Egor V Kostylev, Juan L Reutter, Miguel Romero, and Domagoj Vrgoč. SPARQL with property paths. In *International Semantic Web Conference*, pages 3–18. Springer, 2015.

[77] Laura Krier. Serials, FRBR, and library linked data : A way forward. *Journal of Library Metadata*, 12(2-3) :177–187, 2012.

[78] Javier Lacasta, Javier Nogueras-Iso, Gilles Falquet, Jacques Teller, and F Javier Zarazaga-Soria. Design and evaluation of a semantic enrichment process for bibliographic databases. *Data & Knowledge Engineering*, 88 :94–107, 2013.

[79] Gregory H Leazer and Richard P Smiraglia. Bibliographic families in the library catalog : A qualitative analysis and grounded theory. *Library Resources and Technical Services*, 43(4) :191–212, 1999.

[80] Seungmin Lee and Elin K. Jacob. An Integrated Approach to Metadata Interoperability. *Library Resources & Technical Services*, 55(1) :17–32, 2011.

[81] Seymour Lubetzky. *Principles of cataloging*, volume 1. Institute of Library Research, University of California, 1969.

[82] Hugo Miguel Álvaro Manguinhas, Nuno Miguel Antunes Freire, and José Luis Brinquete Borbinha. FRBRization of MARC records in multiple catalogs. *Joint conference on Digital libraries*, 2010.

[83] Yannis Marketakis, Nikos Minadakis, Haridimos Kondylakis, Konstantina Konsolaki, Georgios Samaritakis, Maria Theodoridou, Giorgos Flouris, and Martin Doerr. X3ML mapping framework for information integration in cultural heritage and beyond. *International Journal on Digital Libraries*, 18(4) :301–319, 2017.

[84] Kristin E Martin, Judith Dzierba, Lynnette Fields, and Sandra K Roe. Consortial cataloging guidelines for electronic resources : I-Share survey and recommendations. *Cataloging & Classification Quarterly*, 49(5) :361–386, 2011.

[85] T P Meehan. What's wrong with MARC ? *Catalogue and Index*, 174 :33–42, 2014.

[86] Hiba MELHEM. *Use and application of semantic web in digital libraries.* Theses, UNIVERSITE GRENOBLE ALPES/ Laboratoire GRESEC (Groupe de recherche sur les enjeux de la communication), 2017.

[87] Pablo N Mendes, Max Jakob, Andrés Garc\'\ia-Silva, and Christian Bizer. DBpedia spotlight : shedding light on the web of documents. In *Proceedings of the 7th international conference on semantic systems*, pages 1–8. ACM, 2011.

[88] Tanja Mercun, Katarina Svab, Viktor Harej, and Maja Zumer. Creating better library information systems : the road to FRBR-land. *Information Research*, 18(3), 2013.

[89] Tanja Merčun and Maja Žumer. New generation of catalogues for the new generation of users : a comparison of six library catalogues. *Program*, 42(3) :243–261, 2008.

[90] Tanja Merčun, Maja Žumer, and Trond Aalberg. Presenting bibliographic families : Designing an FRBR-based prototype using information visualization. *Journal of Documentation*, 72(3) :490–526, 2016.

[91] Erik Mitchell and Carolyn McCallum. Old data, new scheme : An exploration of metadata migration using expert-guided computational techniques. *Proceedings of the American Society for Information Science and Technology*, 49(1) :1–10, 2012.

[92] Ndapandula Nakashole, Gerhard Weikum, and Fabian Suchanek. Discovering and exploring relations on the web. *Proceedings of the VLDB Endowment*, 5(12) :1982–1985, 2012.

[93] Felix Naumann and Melanie Herschel. An introduction to duplicate detection. *Synthesis Lectures on Data Management*, 2(1) :1–87, 2010.

[94] Alireza Noruzi. FRBR and Tillett's Taxonomy of Bibliographic Relationships. *Knowledge organization*, 39(6) :409–416, 2012.

[95] Mark Notess, Jon W Dunn, and Juliet L Hardesty. Scherzo : a FRBR-based music discovery system. In *Proceedings of the 2011 International Conference on Dublin Core and Metadata Applications*, pages 182–183. Dublin Core Metadata Initiative, 2011.

[96] Andrea Giovanni Nuzzolese, Aldo Gangemi, Valentina Presutti, and Paolo Ciancarini. Encyclopedic knowledge patterns from wikipedia links. In *International Semantic Web Conference*, pages 520–536. Springer, 2011.

[97] Thomas Orgel, Martin Höffernig, Werner Bailer, and Silvia Russegger. A metadata model and mapping approach for facilitating access to heterogeneous cultural heritage assets. *International Journal on Digital Libraries*, 15(2-4) :189–207, 2015.

[98] Paul Otlet. *Traité de documentation : le livre sur le livre, théorie et pratique.* Editiones mundaneum, 1934.

[99] Luis Miguel Sintra Salvo Paiva. *Semantic relations extraction from unstructured information for domain ontologies enrichment.* PhD thesis, Universidade Federal do Oeste do Par{á}, 2015.

[100] Shri Ram Pandey, K C Panda, and Others. Semantic solutions for the digital libraries based on semantic web technologies. *Annals of Library and Information Studies (ALIS)*, 61(4) :286–293, 2015.

[101] George Papadakis, George Papastefanatos, and Georgia Koutrika. Supervised meta-blocking. *Proceedings of the VLDB Endowment*, 7(14) :1929–1940, 2014.

[102] M Cristina Pattuelli. Modeling a domain ontology for cultural heritage resources : A user-centered approach. *Journal of the American Society for Information Science and Technology*, 62(2) :314–342, 2011.

[103] Manolis Peponakis. In the Name of the Name : RDF literals, ER attributes and the potential to rethink the structures and visualizations of catalogs. *arXiv preprint arXiv :1609.02004*, 2016.

[104] Manolis Peponakis, Michalis Sfakakis, and Sarantos Kapidakis. FRBRization : using UNIMARC link fields to identify Works. 2011.

[105] Jan Pisanski, Tanja Merčun, and Maja Žumer. FRBR : The Way Forward. *Przegl{\k{a}}d Biblioteczny*, 1(83) :62–72, 2015.

[106] Jan Pisanski and Maja Zumer. User verification of the FRBR conceptual model. *Journal of documentation*, 68(4) :582–592, 2012.

[107] Jan Pisanski, Maja Žumer, and Trond Aalberg. Frbrisation : Towards a Bright New Future for National Bibliographies. In *75th World Library and Information Congress*, 2009.

[108] Michaela Putz, Verena Schaffner, and Wolfram Seidler. FRBR—The MAB2 Perspective. *Cataloging & classification quarterly*, 50(5-7) :387–401, 2012.

[109] Delip Rao, Paul McNamee, and Mark Dredze. Entity Linking : Finding Extracted Entities in a Knowledge Base. In *Multi-source, multilingual information extraction and summarization*, pages 93–115. Springer, 2013.

[110] Mick RIDLEY. Beyond MARC. In *International conference on the principles and future development of AACR*, pages 229–239, 1998.

[111] Jenn Riley. Enhancing interoperability of FRBR-based metadata. In *International Conference on Dublin Core and Metadata Applications*, pages 31–43, 2010.

[112] Pat Riva. Mapping MARC 21 linking entry fields to FRBR and Tillett's taxonomy of bibliographic relationships. *Library resources & technical services*, 48(2) :130, 2004.

[113] Pat Riva, Martin Doerr, and Maja Zumer. FRBRoo : enabling a common view of information from memory institutions. In *World Library and Information Congress : 74th IFLA General Confrence and Council*. Citeseer, 2008.

[114] Pat Riva and Maja Žumer. Introducing the FRBR Library Reference Model. In *Ifla Wilc 2015*, pages 1–7, 2015.

[115] Matthew Rowe, S Sheffield, and United Kingdom. Data . dcs : Converting Legacy Data into Linked Data. *Context*, 10, 2010.

[116] Jodi Schneider. FRBRzing MARC Records with the FRBR display tool, 2008.

[117] Wei Shen, Jianyong Wang, and Jiawei Han. Entity linking with a knowledge base : Issues, techniques, and solutions. *IEEE Transactions on Knowledge and Data Engineering*, 27(2) :443–460, 2015.

[118] Jaeho Shin, Sen Wu, Feiran Wang, Christopher De Sa, Ce Zhang, and Christopher Ré. Incremental knowledge base construction using deepdive. *Proceedings of the VLDB Endowment*, 8(11) :1310–1321, 2015.

[119] Aurélie Signoles, Corinne Bitoun, and Asuncion Valderrama. Implementing FRBR to improve retrieval of in-house information in a medium-sized international institute. *Cataloging & classification quarterly*, 50(5-7) :402–421, 2012.

[120] Agnès Simon, Adrien Di Mascio, Vincent Michel, and Sébastien Peyrard. We grew up together : data. bnf. fr from the BnF and Logilab perspectives. *IFLA 2014*, 2014.

[121] Agnès Simon, Romain Wenz, Vincent Michel, and Adrien Di Mascio. Publishing bibliographic records on the Web of data : Opportunities for the BnF (French National Library). In *Extended Semantic Web Conference*, pages 563–577. Springer, 2013.

[122] Richard P Smiraglia. *Works as entities for information retrieval*. Routledge, 2002.

[123] Elaine Svenonius. 6. Work Languages,". *The Intellectual Foundation of Information Organization*, pages 87–106, 2000.

[124] Patrik Svensson. The landscape of digital humanities. *Digital Humanities*, 2010.

[125] Pedro Szekely, Craig A Knoblock, Fengyu Yang, Xuming Zhu, Eleanor E Fink, Rachel Allen, and Georgina Goodlander. Connecting the smithsonian american art museum to the linked data cloud. In *Extended Semantic Web Conference*, pages 593–607. Springer, 2013.

[126] Mohsen Taheriyan, Craig A Knoblock, Pedro Szekely, and José Luis Ambite. A graph-based approach to learn semantic descriptions of data sources. In *International Semantic Web Conference*, pages 607–623. Springer, 2013.

[127] Naimdjon Takhirov, Trond Aalberg, Fabien Duchateau, and Maja Žumer. FRBR-ML : A FRBR-based framework for semantic interoperability. *Semantic Web*, 3(1) :23–43, 2012.

[128] Naimdjon Takhirov, Fabien Duchateau, and Trond Aalberg. Linking FRBR entities to LOD through semantic matching. In *Theory and Practice of Digital Libraries*. Springer, 2011.

[129] Arlene G Taylor. *Understanding FRBR : what it is and how it will affect our retrieval tools*. Libraries Unltd Inc, 2007.

[130] Teresa Teixeira, Margarida Lopes, Nuno Freire, and B José. Report on FRBR experiments. *Online unter : http ://www. theeuropeanlibrary. org/portal/organisation/cooperation/telplus/documents/TELplus-D3% 206_ v1% 204_ 2008_ 07_ 29. pdf (letzter Zugriff : 22.07. 2011)*, 2008.

[131] Roy Tennant. MARC must die. *Library Journal - New York*, 127(17) :26–27, 2002.

[132] Barbara Tillett. What is FRBR ? A conceptual model for the bibliographic universe. *The Australian Library Journal*, 54(1) :24–30, 2005.

[133] Barbara B Tillett. Bibliographic relationships. In *Relationships in the Organization of Knowledge*, pages 19–35. Springer, 2001.

[134] Yves Tomic. De l'usage des API. *Documentaliste-Sciences de l'Information*, 51(3) :17–18, 2014.

[135] Katerina Tzompanaki and Martin Doerr. A new framework for querying semantic networks. In *Proceedings of Museums and the Web 2012 : the international conference for culture and heritage on-line*, 2012.

[136] Seth Van Hooland, Max De Wilde, Ruben Verborgh, Thomas Steiner, and Rik de Walle. Exploring entity recognition and disambiguation for cultural heritage collections. *Digital Scholarship in the Humanities*, 30(2) :262–279, 2015.

[137] Sherry L Vellucci. Bibliographic relationships. In *The Principles and Future of AACR : Proceedings of the International Conference on the Principles and Future Development of AACR, Toronto, Ontario, Canada, Oct. 23–25*, pages 105–146, 1997.

[138] Ruben Verborgh, Olaf Hartig, Ben De Meester, Gerald Haesendonck, Laurens De Vocht, Miel Vander Sande, Richard Cyganiak, Pieter Colpaert, Erik Mannens, and Rik de Walle. Querying datasets on the web with high availability. In *International Semantic Web Conference*, pages 180–196. Springer, 2014.

[139] Diane Vizine-Goetz. Classify : a FRBR-based research prototype for applying classification numbers. *NextSpace*, 14(2010)(14) :14–15, 2010.

[140] Justyna Walkowska and Marcin Werla. Advanced Automatic Mapping from Flat or Hierarchical Metadata Schemas to a Semantic Web Ontology. In *Theory and Practice of Digital Libraries*, pages 260–272. Springer, 2012.

[141] Shenghui Wang, Antoine Isaac, Stefan Schlobach, Lourens van der Meij, and Balthasar Schopman. Instance-based Semantic Interoperability in the Cultural Heritage. *Semantic Web*, 3(1) :45–64, 2012.

[142] Robert West, Evgeniy Gabrilovich, Kevin Murphy, Shaohua Sun, Rahul Gupta, and Dekang Lin. Knowledge base completion via search-based question answering. In *Proceedings of the 23rd international conference on World wide web*, pages 515–526. ACM, 2014.

[143] Steven Euijong Whang and Hector Garcia-Molina. Joint entity resolution. In *Data Engineering (ICDE), 2012 IEEE 28th International Conference on*, pages 294–305. IEEE, 2012.

[144] Steven Euijong Whang, David Menestrina, Georgia Koutrika, Martin Theobald, and Hector Garcia-Molina. Entity resolution with iterative blocking. In *Proceedings of the 2009 ACM SIGMOD International Conference on Management of data*, pages 219–232. ACM, 2009.

[145] Max De Wilde and Simon Hengchen. Semantic Enrichment of a Multilingual Archive with Linked Open Data. Number JANUARY, 2016.

[146] William E Winkler. The state of record linkage and current research problems. In *Statistical Research Division, US Census Bureau*. Citeseer, 1999.

[147] Annie Wu, Richard Guajardo, and Stephanie Rodriguez. Large-Scale RDA Enrichment of Legacy Data at the University of Houston System Libraries. *Cataloging & Classification Quarterly*, 2016.

[148] Martha M Yee. FRBRization : A method for turning online public finding lists into online public catalogs. *Information technology and libraries*, 24(3), 2005.

[149] Elena I Zagorskaya. Bibliographic relationships in the catalogue, rules and formats. *International cataloguing and bibliographic control*, 29(1) :15–18, 2000.

[150] S. Zapounidou, M. Sfakakis, and C. Papatheodorou. Representing and integrating bibliographic information into the Semantic Web : A comparison of four conceptual models. *Journal of Information Science*, page 0165551516650410, 2016.

[151] Sofia Zapounidou, Michalis Sfakakis, and Christos Papatheodorou. Highlights of library data models in the era of Linked Open Data. In *Research Conference on Metadata and Semantic Research*, pages 396–407. Springer, 2013.

[152] Yin Zhang and Athena Salaba. *Implementing FRBR in Libraries : Key Issues and Future Directions*. Neal-Schuman Publishers, 2009.

[153] Yin Zhang and Athena Salaba. What Do Users Tell Us about FRBR-Based Catalogs ? *Cataloging & Classification Quarterly*, 50(5-7) :705–723, 2012.

[154] Maja Žumer. Functional requirements for bibliographic records : FRBR : The end of the road or a new beginning. *Bulletin of the American Society for Information Science and Technology*, 33(6) :27–29, 2007.

[155] Maja Žumer and Edward T O'Neill. Modeling Aggregates in FRBR. *Cataloging & classification quarterly*, 50(5-7) :456–472, 2012.