Titre: Title:	Génération de base de connaissance à partir de données hétérogènes dans le monde culturel
Auteur: Author:	Dominique Piché
Date:	2022
Type:	Mémoire ou thèse / Dissertation or Thesis
Référence: Citation:	Piché, D. (2022). Génération de base de connaissance à partir de données hétérogènes dans le monde culturel [Mémoire de maîtrise, Polytechnique Montréal]. PolyPublie. https://publications.polymtl.ca/10327/

URL de PolyPublie: PolyPublie URL:	https://publications.polymtl.ca/10327/
Directeurs de recherche: Advisors:	Amal Zouaq, & Michel Gagnon
Programme: Program:	Génie informatique

POLYTECHNIQUE MONTRÉAL
affiliée à l'Université de Montréal

Génération de base de connaissance à partir de données hétérogènes dans le monde culturel

DOMINIQUE PICHÉ
Département de génie informatique et génie logiciel

Mémoire présenté en vue de l'obtention du diplôme de *Maîtrise ès sciences appliquées*
Génie informatique

Mai 2022

POLYTECHNIQUE MONTRÉAL
affiliée à l'Université de Montréal

Ce mémoire intitulé :

Génération de base de connaissance à partir de données hétérogènes dans le monde culturel

présenté par **Dominique PICHÉ**
en vue de l'obtention du diplôme de *Maîtrise ès sciences appliquées*
a été dûment accepté par le jury d'examen constitué de :

Foutse KHOMH, président
Amal ZOUAQ, membre et directrice de recherche
Michel GAGNON, membre et codirecteur de recherche
Frédéric CUPPENS, membre

RÉSUMÉ

Le monde culturel québécois est riche et varié, et ceci se concrétise par l'importante quantité de métadonnées sur les mondes du livre et du cinéma que les acteurs gouvernementaux, académiques et commerciaux ont accumulé. Cependant, ces données sont actuellement en bonne partie indisponibles au public, et sont encodées dans des bases de données dont les modèles, parfois complexes et généralement incompatibles d'une institution à l'autre, rendent l'exploitation difficile. De plus, sauf certaines exceptions, elles ne sont pas reliées aux métadonnées diffusées librement ailleurs sur le web, que ce soit par l'entremise de projets collaboratifs publics tels que Wikidata, ou par des acteurs tels que certaines bibliothèques nationales européennes.

La création de bases de connaissances sous forme de graphes peut permettre la démocratisation de ces métadonnées, en simplifiant leur exploitation et en les liant vers d'autres bases de connaissances existantes. Ce mémoire résume notre travail de création de bases de connaissances pour les mondes du cinéma et de la littérature québécois, en particulier la modélisation de modèles ontologiques et la population des graphes à partir de sources relationnelles.

Nous présentons d'abord une base de connaissances pour le domaine du cinéma québécois, qui utilise un jeu de métadonnées fourni par la Cinémathèque québécoise. À partir de scénarios d'utilisation fournis par des experts du milieu, nous développons un modèle ontologique pour ce domaine, et décrivons la conversion des données sources de leur format original vers la base de connaissances finale.

Ensuite, nous avons modélisé et créé une base de connaissances pour le monde du livre québécois. Encore une fois guidé par des demandes de partenaires du milieu, les données proviennent cette fois de sources variées, ce qui complexifie les tâches de modélisation. La base de connaissances servira de source de données pour un prototype d'application web ainsi qu'un système de question-réponses en langage naturel. L'utilisation de diverses sources nous amène à nous pencher sur la tâche d'alignement d'entités décrites à travers celles-ci, afin de retrouver les mêmes entités du monde réel. La rareté d'identifiants uniques partagés entre sources rend la jointure complexe.

Nous proposons une méthode d'alignement basée sur les modèles de langues neuronaux, une technologie de l'état de l'art du traitement automatique du langage naturel. Seuls deux travaux existants ont attelé ces modèles à cette tâche. Nous vérifions donc si ces techniques seraient également utiles pour l'alignement d'entités du monde de la culture, en langue française.

ABSTRACT

Quebec's cultural world is rich and full of variety, as is illustrated through the imposing amount of cultural heritage metadata that exists. Governmental, academic and commercial players have accumulated a large amount of data relating to literary and film works. However, this data is currently largely unavailable to the greater public, and are held in datastores whose underlying datamodels, which are often complex and incompatible between institutions, complicate their use. On top of this, except for rare exceptions, this data is not interlinked with other linked open data sources available elsewhere on the web, whether they be public collaborative projects such as Wikidata or knowledge bases published by national librairies.

The development of knowledge bases in graph form can aid in democratising this metadata, by simplifying its exploitation and allowing it to be linked with existing, open knowledge bases. This memoir summarizes our work, which is the creation of knowledge bases for Quebec's cinema and literature data. In particular, our work focuses on modelling and populating such knowledge bases from existing relational databases.

We first model a knowledge base for Quebec's film world, which uses a dataset provided by the Cinémathèque québécoise. Our use cases, provided by domain experts, guide our development of an ontological model for this domain. We describe the translation of this source data from its original format towards the final knowledge base.

We then present a knowledge base we've modeled and created for Quebec's world of books. Once again, our modelling is guided by use cases from our domain partners. However, the source data for this knowledge base consists of a collection of heterogenous data sets, which complexifies the modelling task. The resulting knowledge base then serves as a back-end for a web application prototype, as well as a natural language question answering system. The integration of multiple sources brings to light another problem, which we examine: entity alignement between these various sets, to determine which describe the same real-world entities. The scarcity of global unique identifiers makes joining these sets a complex task.

We propose an entity alignment method based on neural langue models, a state of the art natural language processing technique. Only two existing works have previously used this technology for this task. We thus verify if these techniques are also useful for aligning cultural heritage entities in French.

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

LISTE DES SIGLES ET ABRÉVIATIONS

API	Application Programming Interface
BAnQ	Bibliothèque et Archives nationales du Québec
CIDOC	Comité de documentation du Conseil international des musées
CRM	Conceptual Reference Model
FRBR	Function Requirements for Bibliographic Records
FRBR	Function Requirements for Bibliographic Records Object Oriented
GLUE	General Language Understanding Evaluation
IFLA	International Federation of Library Associations
ISBN	International Standard Book Number
ISNI	International Standard Name Identifier
LRM	Library Reference Model
MARC	Machine Readable Cataloguing
MCCQ	Ministère de la Culture et des Communications du Québec
MLM	Masked Language Modeling
OWL	Web Ontology Language
RDA	Resource Description Access
RDF	Resource Description Framework
SPARQL	SPARQL Protocol and RDF Query Language
TAL	Traitement automatique de la langue naturelle
URI	Uniform Resource Identifier
VIAF	Virtual International Authority File
XML	Extensible Markup Language

LISTE DES ANNEXES

CHAPITRE 1 INTRODUCTION

La quantité de données disponibles librement sur le web a connu une croissance fulgurante au cours des dernières années. Cependant, ces ressources sont généralement distribuées à travers une multitude de serveurs indépendants, avec peu ou pas d'interconnexion. Les informations sur un sujet donné se retrouvent fréquemment dispersées sur divers sites, en particulier dans le monde de la culture, constitué d'une multitude d'acteurs indépendants. La tâche d'identifier ces sites et de regrouper l'information requise revient, sur le web "classique", à un humain.

Les données du monde de la culture sont d'un intérêt important pour un grand nombre d'utilisateurs, dont les chercheurs, les écoles, diverses organisations et associations. Leurs cas d'utilisation, incluant l'intégration, l'analyse, l'annotation et la visualisation, sont difficilement réalisables avec des données isolées les unes des autres. Établir des liens facilement navigables par des ordinateurs entre les différentes sources de données est un travail crucial à la facilitation de l'exploitation de ces ressources.

1.1 Éléments de la problématique

Le monde de la culture souffre du manque de navigabilité et de liaison entre silos de données. Les métadonnées des oeuvres culturelles, incluant livres et films, sont fréquemment dispersées. Les acteurs du milieu - bibliothèques nationales, éditeurs et sites communautaires (IMDb, Babelio, etc.) - disposent chacun de données de natures et perspectives différentes sur ces oeuvres. La facilité et le format d'accès à ces données varie grandement selon la mission de l'organisation les détenant. Les institutions culturelles nationales, chargées de missions de préservation, accumulent une quantité imposante de données, mais n'ont pas forcément la mission de rendre celles-ci facilement accessibles et navigables. Les sites communautaires, ayant comme modèle l'interaction avec l'utilisateur, mettent leur données à disposition à travers leurs interfaces web. Les éditeurs et autres entités commerciales publient rarement leurs données, vu qu'elles leur servent principalement à des fins de gestion.

Le web sémantique est une solution fréquemment proposée pour résoudre ce type de problème. Il s'agit d'une sous-discipline de l'intelligence artificielle qui consiste en un ensemble de technologies visant à rendre le web utilisable par la machine. Le web sémantique cherche à permettre à des raisonneurs automatiques de pouvoir effectuer des liens entre ressources, organisées en structure de graphe, grâce à un système de métadonnées formelles. On retrouve notamment parmi ses objectifs la possibilité de développer des outils logiciels pouvant

répondre à des requêtes précises pour lesquelles l'information n'est pas nécessairement stockée de manière explicite, par le biais d'inférences. Un autre objectif est l'interopérabilité de données physiquement stockées sur différents serveurs hôtes par l'entremise de liens, afin d'étendre la portée de moteurs de requêtes au-delà des données stockées dans leur environnement local. L'idée est de permettre à des outils automatiques d'effectuer un cheminement similaire à celui qu'effectue un humain naviguant le web « traditionnel », c'est-à-dire de suivre des liens entre pages web et raisonner sur la base d'informations tirées de pages différentes.

Cette technologie n'est pas, pour l'instant, répandue, mais suscite un intérêt dans des domaines variés, notamment là où différentes sources de connaissances se rapportant à un même domaine sont disponibles, mais difficile d'accès, soit à cause de manque de liens entre sources de connaissances ou par la difficulté d'extraction d'information utile à partir d'une source particulière, comme c'est le cas pour les données culturelles au Québec. Plusieurs aspects nécessitent une attention particulière lors du développement de plateformes de web sémantique, certains indépendants du domaine, certains spécifiques au monde culturel.
Un premier défi est la modélisation de l'ontologie qui structure un graphe de connaissances, qui doit être à la fois simple d'exploitation et suffisamment complexe pour représenter les multiples nuances des aspects du monde que l'on souhaite représenter. Plus un modèle doit contenir d'informations précises, plus il doit être complexe. En revanche, plus il est complexe, plus la charge cognitive peut être élevée pour l'utilisateur, particulièrement si celui-ci n'a pas de connaissance préalable poussée des ontologies ou du domaine décrit par une ontologie donnée. De plus, les requêtes peuvent être complexes et lourdes à exécuter.

La qualité des données sources, parfois bonne, parfois présentant des lacunes sévères, est également une source de problèmes rencontrés lors de la génération d'un graphe de connaissances. Un alignement direct entre le format d'origine et le format du graphe est difficile lorsque les données sont structurées de manière inconsistante à l'intérieur d'une source, puisqu'on ne peut se fier au fait qu'un champ d'origine contienne toujours la même information. Les sources présentent parfois d'autres irrégularités ou particularités, telles que des erreurs de frappe, des inconsistances dans les vocabulaires de catégorisation, l'évolution de standards de catalogage, et ainsi de suite. Avant de pouvoir traduire les données dans un nouveau format, il est donc impératif de repérer la présence, la fréquence et la nature de ce type de problèmes, afin de pouvoir les enrayer ou les éviter.

Un autre défi important se présente lorsque plusieurs sources de données hétérogènes doivent être intégrées à une seule et unique base (ou graphe) de connaissances. Différentes sources

utilisent souvent différents vocabulaires, différents standards de catalogage, et ont des portées et des compréhensions différentes du domaine. Il faut ainsi comparer les modèles et les données sources pour retrouver les similarités et les différences, et en tenir compte lors de la modélisation d'une ontologie servant à conserver un maximum de détail de chacune de ces sources.

Finalement, l'hétérogénéité des données est également problématique lorsque vient le temps d'identifier quelles entités de différentes sources représentent la même entité du monde réel. Cette tâche, l'alignement d'entités, peut facilement se faire en la présence d'identifiants globaux uniques tels que l'ISBN d'un livre. Cependant, un grand nombre d'entités cruciales n'ont pas d'identifiants globaux uniques, par exemple les éditeurs. Il faut donc se replier sur l'utilisation des caractéristiques de ces entités, telles que leurs noms, leurs dates associées, etc., afin de déterminer quelles entités sont identiques. Cette tâche est rendu plus difficile en raison des différences entre les contenus et les portées des champs des différentes sources de donnéese. Bien qu'il s'agisse d'un problème connu, de nouvelles solutions proposées dans les dernières années, dont les modèles de langue neuronaux, pourraient permettre de le résoudre facilement.

1.2 Objectifs de recherche

Le présent mémoire relate le développement de deux projets de développement de plateformes de web sémantique pour le domaine culturel québécois. Le premier projet, mené en collaboration avec la Cinémathèque québécoise, consiste à développer une telle plateforme pour les métadonnées du monde du cinéma québécois à partir des données relationnelles de la collection de la Cinémathèque. Le deuxième projet, effectué avec le Ministère de la Culture Québécois (MCCQ), a un objectif similaire concernant les données du monde du livre ; il diffère cependant par le fait qu'il intègre des données de diverses sources hétérogènes en envergure, en structure et en qualité. L'intégration de ces sources différentes a fait émerger une troisième avenue de recherche : l'alignement automatique d'entités provenant de sources hétérogènes.

Bien qu'il existe des modèles ontologiques existants, tels que ceux présentés à la section 2.2, chacun d'entre eux présente certaines lacunes les rendant insuffisants pour les projets menés. Notre question de recherche principale est donc la suivante :

Comment l'utilisation de technologies de web sémantique peut-elle aider à faciliter l'homogénéisation, la distribution et l'accessibilité des données du monde culturel québécois ? De cette question principale découlent des sous-questions plus précises, motivées par les difficultés énumérées ci-haut.

1. Q1 : Comment modéliser une ontologie pour le domaine de la culture québécoise et traduire en graphe de connaissances des données sous forme relationnelle ?
2. Q2 : Comment modéliser une ontologie permettant d'intégrer le contenu de plusieurs sources hétérogènes de manière homogène, tout en conservant un maximum d'expressivité ?
3. Q3 : Les modèles de langue neuronaux sont-ils un outil approprié pour aligner des entités hétérogènes du monde culturel ?
   (a) Dans le cadre de l'alignement d'entités du monde culturel, comment les modèles de langue préentrainés se comparent-ils aux règles heuristiques ?
   (b) Quel est l'impact du prétraitement et des formats d'entrée sur la performance de l'alignement à l'aide de modèles de langue ?

1.3 Plan du mémoire

Le chapitre 2 offre une explication des concepts de base du web sémantique, des ontologies et de la tâche d'alignement d'entités, ainsi qu'une revue des ontologies pertinentes existantes et des travaux précédents concernant l'alignement grâce aux modèles de langue.

Le chapitre 3 montre le travail de modélisation et implémentation d'une ontologie pour les données du monde du cinéma québécois (CineTV), à partir d'une unique source de données homogène. On y présente la méthodologie utilisée, et les requis pour la base de connaissances finale, le modèle final retenu ainsi que la vérification par divers moyens des fonctionnalités et contenu du graphe final.

Le chapitre 4 illustre le cheminement effectué pour développer une base de connaissances du monde du livre québécois (MCCQ), cette fois-ci à partir de données provenant de sources multiples et hétérogènes. Bien que l'objectif final soit similaire, beaucoup de complexité vient s'ajouter de par le fait de cette hétérogénéité, ce qui est expliqué notamment dans les sous-sections du chapitre ayant trait à l'analyse des données initiales et au processus de traduction des formats source vers le graphe final.

Au chapitre 5, l'étape d'alignement d'entités nécessaire à la réalisation d'une base de connaissances (chapitre 4) est élaborée en détail ; on y présente un travail d'analyse des capacités d'un modèle de langue à aligner des données de différents niveaux d'hétérogénéité.

Au chapitre 6, nous résumons notre travail, les conclusions que nous en tirons et présentons des pistes d'amélioration.

CHAPITRE 2 CONCEPTS DE BASE ET REVUE DE LITTÉRATURE

Le présent chapitre relate les concepts fondamentaux nécessaires à la compréhension des chapitres suivants, ainsi qu'un résumé de la littérature existante.

Dans la section 2.1, nous introduisons les concepts du web sémantique. Nous y abordons plus particulièrement les graphes de connaissances, la structure selon laquelle les connaissances sont organisées, et les ontologies, constituées d'axiomes logiques permettant de définir les connaissances sous-jacentes à ces graphes.

La section 2.2 présente les ontologies et taxonomies existantes du monde culturel qui pourraient servir de base à une ontologie du monde de la culture québécoise.

Finalement, la section 2.3 explique le concept d'alignement d'entités et de modèles de langue, et présente les travaux existants utilisant des modèles de langue pour l'alignement.

2.1 Web sémantique et graphes de connaissances

Le web sémantique est une extension du World Wide Web qui a pour but de rendre les données du web lisibles par la machine. Pour ce faire, il est nécessaire d'encoder non seulement les données en elles-mêmes, dans un graphe de connaissances, mais également de l'information sémantique permettant de spécifier les liens entre différentes entités, par l'entremise d'une ontologie. Ces concepts sont expliqués dans les paragraphes suivants.

2.1.1 Graphe de connaissances

Un graphe de connaissances KG est un encodage de données sous la forme d'un graphe, constitué de noeuds et d'arêtes directionnelles. Chaque noeud représente une entité E, et chaque arête représente une relation R entre deux noeuds. Ainsi, l'ensemble d'un graphe de connaissances peut être décrit selon les relations entre noeuds et arêtes qu'ils contient :

$$KG \subseteq E \times R \times E$$

RDF [4] (*Resource Description Framework*) est un cadre de structure syntaxique pour la représentation d'information sur le web sous forme de graphes de connaissances. Un graphe RDF peut être décrit comme étant constitué d'une suite de triplets de la forme (s, p, o), où s est le sujet, dont la relation à l'objet o est le prédicat p. Un graphe RDF peut être encodé

selon une de plusieurs syntaxes concrètes, dont la syntaxte Turtle [1], qui sera utilisée pour le reste du présent mémoire.

Chacun de ces trois éléments est représenté par un identifiant global unique (URI – *Uniform Resource Identifier*), à l'exception de l'objet, qui peut être soit une URI soit une valeur telle qu'un nombre ou une chaîne de caractères.

Dans l'exemple suivant, exprimé en sérialisation Turtle d'un triplet RDF, on a comme sujet une personne (Jean Dujardin), comme prédicat la relation d'une personne à sa date de naissance, et comme objet la valeur de cette date pour le sujet en question (le 19 juin 1972).

```
example:JeanDujardin schema:birthDate "1972-06-19"^^xsd:date .
```

Dans l'exemple précédent, l'objet contient une valeur et non un URI, faisant du prédicat une *DatatypeProperty*. Par contre, si on voulait décrire la relation entre deux entités, tel que la relation entre un acteur et un film dans lequel il a joué, on pourrait avoir comme sujet et objet les URI des entités représentant ces deux concepts, la personne et le film, et comme prédicat l'URI de l'action d'avoir joué dans un film. Dans ce cas, le prédicat est une *ObjectProperty*. Le prochain exemple consiste en deux triplets RDF liant des ressources. Le premier exprime la relation de mariage entre les personnes Jean Dujardin et Nathalie Péchalat, et le deuxième, le fait que Jean Dujardin est de nationalité française.

```
example:JeanDujardin schema:spouse example:NathaliePechalat ;
                     schema:nationality dbp:France .
```

L'exemple précédent illustre également un des raccourcis syntaxiques disponibles en Turtle : comme les deux triplets ont le sujet Jean Dujardin, il n'est pas nécessaire de répéter son identifiant deux fois. De plus, il montre l'utilisation d'espaces de noms permis par la syntaxe RDF. Il est possible de déclarer des préfixes que l'on peut employer pour simplifier l'écriture de triplets. Par exemple, la déclaration suivante du préfix *schema* :

```
@prefix schema: <https://schema.org/> .
```

permet de remplacer l'URI `<https://schema.org/nationality>` par `schema:nationality`.

Les **données liées** [5] prennent la forme de graphes de connaissances, avec une spécification importante : chaque identificateur de ressource est unique non seulement dans le graphe de connaissances local, mais également universellement. Il est donc possible d'utiliser des URI

1. https ://www.w3.org/TR/turtle/

pour des entités ou des relations qui ne sont pas explicitement décrites dans le graphe local, mais qui le sont sur un serveur distant. Ainsi, un moteur de recherche peut fédérer une partie d'une requête à un autre serveur hôte, combinant ainsi les informations locales et distantes dans une seule et même réponse à l'utilisateur, de manière transparente pour celui-ci.

Prenons deux serveurs contenant les données respectives des figures 2.1 et 2.2.

```
@prefix : <http://example.org/> .

:acteur8126 :nom :"Nathalie Péchalat"^^xsd:string .
:acteur1023 :nom :"Jean Dujardin"^^xsd:string .
```

Figure 2.1 Données contenues sur le serveur source1.example.org

```
@prefix : <http://example.org/> .
@prefix schema: <https://schema.org/> .

:acteur8126 schema:spouse :acteur1023 .
```

Figure 2.2 Données contenues sur le serveur source2.example.org

La requête 2.3 interroge les deux serveurs, le premier contenant des noms d'individus, et le deuxième des relations entre personnes. Elle retournera le nom de la personne mariée à la personne s'appelant "Nathalie Péchalat", c'est-à-dire la chaîne de caractères "Jean Dujardin".

```
PREFIX schema:   <http://schema.org/>
SELECT ?nomPersonne2
WHERE
{
    SERVICE <http://source1.example.org/sparql> {
        ?personne1 :nom "Nathalie Péchalat"^^xsd:string .
        ?personne2 :nom ?nomPersonne2 .
        SERVICE <http://source2.example.org/sparql> {
            ?personne1 schema:spouse ?personne2 .
        }
    }
}
```

Figure 2.3 Requête fédérée à deux modules de requêtes distants

C'est à travers ce mécanisme que le web sémantique permet d'inter-relier des données hébergées sur des serveurs distants.

Ontologies Une ontologie est un ensemble de concepts et de règles sémantiques les reliant. Ceci forme un modèle régissant la structure que peut prendre un ensemble de données. Par exemple, une ontologie sur des oeuvres littéraires pourrait comprendre des concepts comme "Personne", "Livre", "Éditeur", "Endroit de publication" et des règles, telles que "Un livre peut être publié à un ou plusieurs endroits", "Un livre doit avoir au minimum un auteur", "Un livre peut avoir une date de publication". L'interaction entre ces éléments impose une structure aux données. Par exemple, la date de naissance d'une personne doit être une date, et ne peut être un lieu. Ces contraintes visent à préserver une cohérence sémantique et à modéliser un aspect de la réalité concernée par l'ontologie. Ces règles sémantiques sont formalisées à l'aide de classes et de propriétés. Dans cet exemple, les classes seraient "Personne", "Éditeur", etc., et une propriété serait "a écrit" le livre, permettant de relier un auteur à un livre.

Lorsque des données sont structurées de cette façon, il est souvent possible d'inférer de nouvelles connaissances. Par exemple, prenons la propriété transitive "est adapté de". On sait qu'un livre A est une adaptation d'un livre B, et que le livre B est une adaptation du livre C. On peut alors inférer que le livre A est une adaptation du livre C, comme illustré à la figure 2.4.

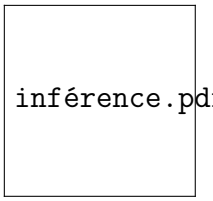

Figure 2.4 Inférence d'une relation à l'aide de propriétés transitives

Une ontologie formalise une certaine vision de la réalité, souvent centrée sur un certain domaine. La réalité étant complexe et nuancée, le niveau de détail et le contenu d'une ontologie doit être adaptée aux besoins. Par exemple, une ontologie traitant de géologie n'a pas besoin de représenter de façon très détaillée, voir du tout, les relations entre individus, contrairement à une ontologie traitant de structures organisationnelles. Il existe une pléthore d'ontologies, allant de très générales, par exemple DBpedia [2], à hautement spécialisées, par exemple la *Gene Ontology* [6], qui traite de génétique.

2. https ://www.dbpedia.org/

Dans la section suivante, nous introduisons diverses ontologies utilisées dans le monde culturel, des plus généralistes aux plus spécifiques.

2.2 Ontologies pour les données culturelles ouvertes

Dans cette section, nous présentons les ontologies et vocabulaires existants pouvant servir de base à des modèles de données ontologiques du monde de la culture québécoise, et motivons les choix effectués.

2.2.1 Ontologies et vocabulaires généralistes

DBpedia

Les utilisateurs de dbpedia.org [7] maintiennent une ontologie généraliste très large (plus de 775 classes et 2861 propriétés) et souvent réutilisée (intégrée à 16 autres vocabulaires, selon le site Linked Open Vocabularies [3]). Cette ontologie n'a pas de domaine spécifique ; elle s'intéresse aux informations pouvant être extraites de Wikipédia. Mis à part les entités principales nécessaires pour la représentation de nos données (œuvre cinématographique, oeuvre littéraire, personne), les relations nécessaires pour représenter toutes les fonctions qu'occupent les personnes dans la production de films ou l'écriture et la publication d'oeuvres littéraires dans le monde du livre n'y sont pas toutes présentes ; seules les relations telles que réalisation, direction photo, écriture de texte et composition de musiques sont présentes.

Schema.org

Comme DBpedia, le vocabulaire de schema.org [8] est généraliste ; il n'a pas été élaboré pour des fins de bibliothéconomie ou de représentation de collections cinématographiques. Cependant, son intérêt principal réside dans son adoption : selon Linked Open Vocabularies-slov, 65 vocabulaires le réutilisent. De ce fait, on retrouve les mêmes limitations que pour l'ontologie de DBpedia : comme il n'a été développé ni pour les domaines littéraire ni celui du cinéma, de nombreuses relations nécessaires pour représenter les fonctions que peuvent occuper les personnes dans la création d'une oeuvre cinématographique ou littéraire sont asbentes.

3. https ://lov.linkeddata.es/dataset/lov/

Dublin core

Le Dublin Core Schema [9] est un vocabulaire restreint qui sert à décrire des ressources digitales et physiques. Il s'agit du vocabulaire le plus réutilisé dans le monde, avec 590 vocabulaires l'intégrant (toujours selon Linked Open Vocabularieslov. Bien que ce vocabulaire comporte un grand nombre d'entités intéressantes quant aux métadonnées globales d'une œuvre, le concept même d'un film est absent. Ce vocabulaire est donc de trop haut niveau.

2.2.2 Ontologies et vocabulaires bibliographiques et des arts

Art and Architecture Thesaurus (AAT)

Le AAT [10] est un vocabulaire contrôlé de Getty utilisé pour décrire les entités ayant rapport à l'art, à l'architecture et à la culture matérielle. Il est utilisé par des musées, des librairies d'art, des archives, entre autres. Bien que quelques termes des mondes littéraires et du cinéma s'y retrouvent, tel que *auteur* ou *producteur de film*, la grande majorité du vocabulaire se concentre sur la description d'entités physiques telles que des églises, des toiles, etc. De plus, de nombreuses fonctions que peuvent occuper les personnes participant à la création et la publication d'oeuvres cinématographiques et littéraires ne s'y retrouvent pas.

FRBR (Functional Requirements for Bibliographic Records)

Les spécifications fonctionnelles des notices bibliographiques (FRBR) [11] sont une modélisation entité-relation, développée par la Fédération internationale des associations et institutions de bibliothèques (IFLA). Les notices bibliographiques forment une partie importante des métadonnées disponibles sur le monde littéraire, et FRBR a été conçu afin de représenter celles-ci de manière flexible et complète. FRBR, indépendant de tout code de catalogage ou implémentation, représente une vue générale du domaine de la bibliothéconomie. Ce modèle permet de représenter trois groupes principaux d'entités : les œuvres, les acteurs et les sujets d'œuvre. Les œuvres sont déclinées en quatre entités plus spécifiques, représentant différents niveaux d'abstraction : œuvres, expressions, manifestations et items. Ce modèle permet la description à la fois des différentes versions conceptuelles d'une œuvre (adaptations régionales, par exemple) et des versions physiques (différentes copies). Cependant, la version entité-relation de FRBR n'est pas adaptée à la représentation en base de connaissances.

RDA (Resource Description and Access)

RDA : Resource Description and Access [4] [5] est un ensemble d'outils de données liées utilisé pour le catalogage de données bibliographiques. Adopté par les bibliothèques nationales des États-Unis, du Canada, de la Grande Bretagne, de l'Australie et de l'Allemagne, RDA est un standard assez répandu pour rendre l'interopérabilité entre les données de la Cinémathèque et du MCCQ et les ontologies l'utilisant intéressante. De plus, RDA respecte FRBR (Functional Requirements for Bibliographic Records). Comme DBpedia et schema.org, RDA ne comprend pas suffisamment de types d'entités pour représenter adéquatement les fonctions qu'occupent les personnes dans la production de films ou l'écriture et la publication d'oeuvres littéraires. On y retrouve les rôles principaux (tels que cinéaste, auteur, réalisateur, directeur de photographie, acteur et éditeur), mais la majorité des rôles sont absents. De plus, l'accès aux outils de développement (RDA Toolkit) est payant.

CIDOC-CRM

CIDOC-CRM [12] (CIDOC : comité de documentation du Conseil international des musées, CRM : Conceptual Reference Model) propose une modélisation de très haut niveau. Le but est d'offrir une structure complète pour la description d'événements et d'entités d'une collection muséale. Une façon d'exploiter CIDOC-CRM est d'implémenter des sous-classes et des sous-propriétés de ces concepts se rapportant directement au domaine à l'étude. L'avantage de l'utilisation de CIDOC-CRM est donc la compatibilité de la modélisation des collections de la Cinémathèque et des données du MCCQ avec d'autres ontologies de collections.

L'approche de l'ontologie CIDOC-CRM est de spécifier les événements temporels comme des entités à part, qui ont lieu pendant des laps spécifiques de temps, et auxquels participent des acteurs de manières diverses. À partir de nos sources de données, on peut déduire l'existence de types d'événements pouvant être des sous-classes du concept d'événement utilisé dans CIDOC-CRM : la publication d'un livre, la naissance ou le décès d'un acteur, la sortie et la production d'un film. On remarque qu'une certaine complexité est apportée dans la représentation d'événements par CIDOC-CRM, puisque plusieurs entités sont nécessaires pour représenter les notions temporelles, au lieu d'une simple valeur. Par exemple, les dates associées à un concept tel que la publication d'un film ne sont plus directement associées au film, mais sont associées à l'entité de laps de temps (*time-span*) associée à l'entité représentenant l'événement de publication. À priori, CIDOC-CRM n'est pas compatible avec le modèle FRBR. Une extension de CIDOC-CRM a été développée pour combler ce manque :

4. http ://www.rda-rsc.org/
5. https ://www.rdatoolkit.org/index.php/

FRBRoo.

FRBRoo

FRBRoo [13] (FRBR object-oriented) est une ontologie issue d'une collaboration entre les groupes de travail de CIDOC-CRM et FRBR, née d'un désir d'interopérabilité des données bilbliographiques et muséales. Ainsi, FRBRoo est dérivé de CIDOC-CRM, et ajoute des classes et des propriétés servant à décrire plus précisément les œuvres et les événements associés à la création de celles-ci. Par exemple, FRBRoo reprend de FRBR les concepts d'œuvre, manifestation et item, qui sont cruciaux pour la représentation de collections en bibliothèque. Ainsi, non seulement les données de CineTV et du MCCQ peuvent être représentées en données liées selon les standards prescrits par FRBR, mais elles peuvent également être exploitées selon la perspective d'une collection de musée (soit CIDOC-CRM). De plus, comme il s'agit d'une ontologie de plus haut niveau et non d'un vocabulaire fixe, il est non seulement encouragé, mais attendu, de créer une implémentation spécifique des concepts plus pointus requis pour la représentation des fonctions dans nos ontologies.

Linked Art

Linked Art [6] [14] est une initiative implémentant le modèle de données CIDOC-CRM [12] et les vocabulaires Getty (dont AAT) en RDF. Bien que cela soit un exemple très illustratif d'une implémentation robuste de CIDOC-CRM, l'utilisation des vocabulaires Getty, qui, comme vu plus haut, ne sont pas suffisamment adaptés aux mondes du cinéma et de la littérature, ainsi que la portée principale du modèle (semblable à celle de AAT) font en sorte que la réutilisation de cette implémentation complexifierait le modèle final, sans fournir tous les outils nécessaires pour la représentation adéquate des données de MCCQ ou CineTV.

ArCo (Architecture of Knowledge)

ArCo [15] est une ontologie développée par le ministère italien de l'héritage et des activités culturelles (MiBAC), et est publiée en RDF. Cependant, ArCo exclut les collections bibliographiques, ce qui a comme effet l'incompatibilité avec des modèles de catalogage répandus tels que FRBR. Cette ontologie respecte plutôt la Normativa Trasversale (Norme Transversale), qui est un encodage standardisé pour le catalogage développé par le MiBAC. Ainsi, l'interopérabilité avec d'autres données ouvertes sera difficile à intégrer à un modèle basé sur ArCo.

6. https ://linked.art/model/

2.2.3 Ontologies du cinéma et du monde littéraire

Creative Works Ontology

L'ontologie Creative Works [16], développée par MovieLabs [7], s'intéresse spécifiquement au domaine du cinéma. Malgré sa grande complexité, cette ontologie n'est pas plus riche en termes de fonctions que DBpedia, schema.org ou RDA, et n'est pas intégrée à d'autres vocabulaires. De plus, la documentation disponible en ligne est éparse. Le seul avantage par rapport aux vocabulaires précédents est l'intégration d'une notion d'événement, ce qui permet une riche représentation d'un événement de sortie d'un film, ou la production de celui-ci. Par contre, les deux ontologies précédentes non seulement règlent le même problème, mais sont mieux adaptées à la bibliothéconomie (et peuvent donc plus complètement représenter les données provenant de catalogues), et sont plus largement adoptées.

IFLA-LRM

Le IFLA Library Reference Model (LRM) [17], paru en 2017, est issu d'un travail de consolidation par l'IFLA des modèles de métadonnées bibliographiques précédents, incluant FRBR. De ce fait, LRM rend FRBR obsolète. IFLA-LRM est une ontologie du monde littéraire.

On remarque notamment la reprise à partir de FRBR des œuvres, expressions, manifestations et items. De plus, les appellations, les laps de temps et les places sont gérées de manière expressive, comme dans CIDOC-CRM, ce qui permet d'encoder avec précision les étendues temporelles et les divers types de noms, d'identifiants et de classificateurs qu'on retrouve dans les systèmes de catalogage. Ce modèle permet également d'intégrer des métadonnées de catalogage des bases de données originales qui seront rendues désuètes par ce modèle, mais dont il serait intéressant de garder des traces.

2.2.4 Modèles retenus

Parmi ces modèles, certains présentent des caractéristiques plus adaptées aux problèmes qui nous intéressent, soit la possibilité de modéliser l'ensemble des données sources, l'interopérabilité avec d'autres bases de connaissances du milieu culturel et la simplicité d'utilisation. Les modèles les plus répandus dans le monde culturel sont CIDOC-CRM, ainsi que les modèles d'IFLA, soit FRBR et son évolution, LRM. Ces modèles sont pleinement en mesure de représenter l'ensemble des concepts et des subtilités des mondes du cinéma et de la lit-

7. https ://movielabs.com/

térature. Par contre, ils peuvent être plus difficiles d'utilisation à cause de leur complexité, qui provient principalement de l'utilisation d'entités représentant des événements ainsi que de la décomposition d'oeuvres en plusieurs entités de niveaux d'abstraction différente. Cette complexité rend plus compliquée l'écriture de requêtes avec ces modèles.

Ainsi, la solution qui sera adoptée sera l'utilisation de deux modèles parallèles. Un modèle plus expressif mais complexe, et donc potentiellement difficile d'utilisation, ainsi qu'un modèle plus simple employant une structure minimaliste et un vocabulaire répandu.

Les modèles expressifs retenus sont ainsi FRBR et LRM, qui sont mieux adaptés à la représentation de données issues de notice bibliographiques que CIDOC-CRM. Quant au modèle simple, il utilise le vocabulaire schema.org, car celui-ci est parmi les vocabulaires les plus réutilisés dans le monde des données liées.

La sous-section 2.2.5 présente une méthodologie de développement d'ontologies qui nous permettra de cerner quels éléments de ces ontologies existantes sont nécessaires pour la création de bases de connaissances pour les mondes littéraires et cinématographiques.

Une fois les modèles ontologiques implémentés, le problème suivant survient lors de la traduction des données vers le graphe de connaissances, plus particulièrement dans le contexte de l'intégration de données de plusieurs sources vers un graphe de connaissances unifié : l'alignement d'entités. À la section 2.3 suivante, cette tâche et les solutions envisagées sont présentées.

2.2.5 Développement de modèles ontologiques

Les modèles existants retenus à la section 2.2.4 comportent la majorité des concepts nécessaires pour décrire les domaines du cinéma et de la littérature. Par contre, il se peut que certains éléments nécessaires soit absents. De plus, ces modèles comportent de nombreux concepts et relations qui ne sont pas nécessaires pour les bases de connaissances que nous développons. Par exemple, IFLA LRM décline le concept de livre en quatre niveaux d'abstractions : l'oeuvre conceptuelle, son expression en langage, la manifestation d'une expression dans un format de distribution, et l'item, qui représente un exemplaire d'une manifestation. Dans un cas où une ontologie n'a pas besoin d'inventorier des copies de livres, il n'est pas nécessaire d'intégrer la classe d'item au modèle.

SAMOD La méthodologie SAMOD (*Simplified Agile Methodology for Ontology Development*) [18] propose de développer des ontologies avec un processus itératif. En résumé, SAMOD propose de procéder de la manière suivante :

1. Des experts du domaine et les développeurs de l'ontologie créent des scénarios motivants, qui décrivent en langage naturel un aspect du domaine, avec des exemples qui illustrent cet aspect et un glossaire des termes importants du domaine du scénario.
2. Chaque scénario motivant est traduit en une ou plusieurs questions de compétence informelles en langage naturel. Chaque question de compétence est accompagnée du type de réponse souhaité, d'exemples de réponses (qui correspondent aux exemples du scénario motivant) et d'une liste de questions de plus haut niveau qui requièrent la présente, s'il y a lieu.
3. Un *modelet* pour le domaine décrit par un scénario est développé. Un modelet est limité aux entités nécessaires pour la description d'un scénario donné et les questions de compétence s'y rattachant. En plus de la définition formelle du modelet ($TBox$), un jeu de données ($ABox$) représentant les exemples du scénario est créé.
4. Les questions de compétence informelles sont traduites en questions de compétence formelles sous la forme de requêtes en SPARQL.
5. Trois vérifications sont alors effectuées pour déterminer si le modelet est valide :
   — Un test du modèle examine le scénario motivant, son glossaire et la définition du modelet ($TBox$) de manière qualitative pour déterminer si le modèle est logiquement valide et s'il couvre adéquatement le domaine du scénario.
   — Un test de données détermine si les contenus de la $TBox$, soit la définition du modelet, et de la $ABox$, soit les données correspondant aux exemples, sont valides. Il vérifie si les données de la $ABox$ sont compatibles avec le modèle de la $TBox$ et si le contenu de la $ABox$ couvre adéquatement les exemples du scénario motivant.
   — Un test de requête vérifie si les questions de compétence formelles en SPARQL s'exécutent correctement sur la combinaison des $TBox$ et $ABox$. Il détermine également si le retour de chaque requête SPARQL se conforme à ce qui est attendu dans la définition informelle de la question de compétence correspondante.
6. Si les tests sont réussis, le modelet est fusionné au modèle existant, si d'autres itérations ont été effectuées précédemment.
7. Les cas de test utilisés sont ajoutés aux cas de tests du modèle existant. Ces cas devront continuer à être réussis si de nouveaux modelets sont rajoutés.

Cette méthodologie bâtit donc progressivement une ontologie pour un domaine à l'aide de scénarios élaborés à l'aide d'experts du domaine. Elle s'assure de la validité et fonctionnalité de l'ontologie. De plus, elle permet d'éviter d'intégrer à l'ontologie des concepts qui ne sont pas nécessaires pour les cas d'utilisation.

2.3 Alignement d'entités

La tâche d'alignement consiste tout simplement à déterminer si deux entités représentent la même entité du monde réel. Unifier en une seule base de connaissances plusieurs sources de données nécessite donc d'aligner les entités de ces sources différentes, pour déterminer quels enregistrements concernent une même et unique entité du monde réel. Lorsque des identifiants uniques sont présents à travers les collections sources, tels que des ISBN dans le cas de livres, cette tâche est triviale, puisqu'il suffit de déterminer quels enregistrements possèdent un même identifiant.

En l'absence de tels identifiants, il est nécessaire d'examiner les caractéristiques des entités, afin de déterminer la valeur des attributs des entités. Par exemple, on peut affirmer avec quasi-certitude que deux auteurs ayant un nom identique, une date et un lieu de naissance identiques, et ayant écrit un livre avec un titre identique sont la même personne.

Cependant, ce genre d'alignement nécessite l'élaboration d'une multitude de règles heuristiques, avec un effort important de développement. De plus, des sources de données hétérogènes peuvent utiliser des standards de formatage de données différents. Par exemple, une source peut spécifier le nom de personnes sous la forme "Nom de famille, Prénom", alors qu'une autre pourrait exprimer la même information sous la forme "Prénom Nom de Famille". Adapter les règles heuristiques à tous les scénarios similaires entre ensembles, et nettoyer préalablement les données pour éviter d'emblée ce genre de scénario, est coûteux en termes de développement.

Ceci motive l'utilisation de techniques plus avancées issues du domaine de l'intelligence artificielle pour résoudre cette tâche de manière plus automatique. Les modèles de langue, qui sont présentés à la sous-section suivante, ont déjà été utilisés pour cette tâche dans quelques travaux par le passé, qui sont résumés à la sous-section subséquente.

2.3.1 Modèles de langue

Les **transformeurs** sont un type de modèle d'apprentissage profond qui servent à effectuer des tâches telles que la traduction ou la génération de résumés, à partir des séquences de données en entrée, par exemple du langage naturel. Dotés d'un mécanisme d'auto-attention leur permettant de traiter chaque portion d'une séquence dans le contexte de l'entiereté du reste de la séquence, ils ont, dans les dernières années, révolutionné le domaine du traitement automatique de la langue naturelle (TAL) [19].

Les **modèles de langue préentraînes** à la BERT [20] (*Bidirectional Encoder Representations from Transformers*) emploient une architecture basée sur les transformeurs. Ils sont

préentrainés sur deux tâches, leur permettant ainsi d'apprendre des plongements contextuels pour les mots. La première tâche consiste en la modélisation de langage masqué. Pour cette tâche, le modèle prend en entrée des séquences dont 15% des mots sont masqués, et il doit prédire ceux-ci. La seconde tâche est la prédiction de phrase suivante, où le modèle doit prédire la probabilité qu'une séquence donnée puisse suivre une séquence précédente. Ces deux tâches sont effectués sur d'immenses corpus, et sont dispendieuses en temps et ressources computationnelles. Une fois ce pré-entraînement complété, les modèles peuvent être réutilisés, après un entraînement plus bref mais plus spécifique, pour des tâches telles que la classification de séquences, l'annotation de séquences ou la réponse à des questions en langue naturelle. Ce concept de réutilisation de modèles pré-entraînés est désigné par l'apprentissage par transfert.

BERT a été suivi par d'autres modèles similaires, mais avec des modifications importantes. Par exemple, DistilBERT [2] comporte 66 millions de neurones, contre 110 millions pour BERT. Son préentraînement consiste en une distillation du modèle BERT original. La distillation consiste à apprendre à un réseau neuronal "étudiant" à reproduire les sorties d'un modèle "enseignant" [21], en l'occurrence BERT. Il retient 97% de la performance de BERT sur l'ensemble de référence GLUE [8]. RoBERTa [22], un autre modèle basé sur BERT, a une architecture identique à ce dernier, mais les corpus d'entraînement et la méthodologie d'entraînement ont été revus. Notamment, 10 fois plus de données sont utilisées pour le préentraînement, et ce dernier n'est effectué que sur la tâche de modélisation de langage masqué. En conséquence, le préentraînement est environ 4 fois plus lent que celui de BERT, alors que les performances sont de 2% à 20% plus élevées [22] sur les tâches GLUE.

Les modèles de langue ci-haut sont entraînés strictement sur des corpus anglais. Ces mêmes architectures ont toutefois étés réutilisées avec, entre autres, des corpus multilingues ou encore francophones, ce qui pourrait être d'intérêt, puisque les données de la culture québécoise sont majoritairement en français. Deux modèles de langue français employant l'architecture de RoBERTa ressortent : CamemBERT [23] et FlauBERT [1].

CamemBERT, le premier modèle monolingue français basé sur RoBERTa, est entraîné sur la portion française du corpus OSCAR, une version préfiltrée de Common Crawl [9] contenant 138 GB de données.

FlauBERT est similaire à CamemBERT sur certains aspects ; basé sur RoBERTa, il reprend son architecture et sa technique d'entraînement, à savoir l'entraînement sur uniquement la tâche de modélisation du langage masqué, sans prédiction de prochaine phrase. Il

8. https ://gluebenchmark.com/
9. https ://commoncrawl.org/

diffère de CamemBERT par le corpus de préentraînement, la stratégie de masquage pour le préentraînement et le segmenteur employés. En contraste avec le segmenteur SentencePiece qu'utilise CamemBERT, FlauBERT utilise un segmenteur de type BPE (*Byte-Pair Encoding*). La stratégie de préentrainement de FlauBERT utilise du masquage de sous-mots, plutôt que de mots entiers comme pour CamemBERT. Finalement, l'ensemble de préentraînement de FlauBERT, qui fait 71 GB de taille, est constitué de 24 sous-corpus, incluant notamment des ensembles partagés dans le cadre de la conférence WMT19 [10], des portions de la collection OPUS [11] et des corpus issus de la collection Wikimedia [12].

Les modèles de langage retenus et leurs principales caractéristiques sont résumés au tableau 2.1. Les modèles précédés par "EN" sont entraînés sur des données anglaises, et les modèles précédés par "M" sur plusieurs langues. La colonne "N Params" représente le nombre de coefficients pour chaque modèle.

Tableau 2.1 Comparaison de modèles de langue [1–3]

Modèle	Couches	N Params	Langue	Corpus	Tokenizer
EN-BERT	12	110 M	En	16 GB	WordP. (30K)
M-BERT	12	110 M	104 lan.	Wikipedia	WordP. (30K)
M-DistilBERT	6	66 M	104 lan.	16 GB	WordP. (30K)
CamemBERT	12	110 M	Fr	138 GB	SentenceP. (32K)
FlauBERT	12	138 M	Fr	71 GB	BPE (50K)

2.3.2 Alignement à l'aide de modèles de langue

Des techniques issues du domaine du traitement automatique de la langue naturelle (TAL) employant des transformeurs pré-entrainés sur des tâches de modélisation de langage masqué (*masked-language modeling*, ou MLM), et donc se conformant au principe de transfert de connaissances, ont démontré des performances du niveau de l'état de l'art sur des tâches en aval reliées à la langue. Comme la tâche d'alignement d'entités peut être reformulée en une classification de deux séquences comme étant identiques ou non, ces modèles peuvent facilement être affinés pour cette tâche. Deux travaux existants [24, 25] utilisant les modèles de langue pour l'alignement d'entités ont déjà démontré des performances plus élevées que les modèles existants sur 13 jeux de données de référence, élaborés dans le cadre de l'évaluation de l'outil DeepMatcher [26]. Le deuxième de ces travaux, Ditto [25], utilise des techniques d'augmentation et d'annotation de données afin de forcer le modèle à apprendre les nuances

10. statmt.org/wmt19/
11. https ://opus.nlpl.eu/
12. https ://meta.wikimedia.org/wiki/Data$_d$*umps*

possibles dans la génération de plongements. Les deux méthodes emploient une couche de neurones pleinement connectée et une couche de classification SoftMax pour la prédiction d'alignement, le tout à la sortie d'un modèle de langue anglais préentraîné (BERT [20], DistilBERT [2], RoBERTA [22] ou XLNet [27]).

Comme ces travaux emploient des jeux de données anglais et que les données source du monde culturel québécois sont généralement en français, le remplacement du modèle anglais dans l'architecture du modèle complet d'alignement avec un modèle de langue multilingue ou français pourrait avoir un impact sur la performance. Les deux MLM RoBERTA et CamemBERT, précédemment décrits, utilisent l'architecture de RoBERTa mais sont préentrainés sur des corpus français, et donc sont particulièrement d'intérêt.

Dans les chapitres 3 et 4, nous décrivons comment les modèles ontologiques retenus précédemment dans ce chapitre sont implémentés, modifiés et étendus pour servir de structure à des bases de connaissances pour les milieux québécois du film et de la littérature. Au chapitre 5, les modèles de langue sont utilisés pour l'alignement d'entités dans le cadre de la création de la base de connaissances du monde du livre.

CHAPITRE 3 CONVERSION D'UNE BASE DE DONNÉES RELATIONNELLE EN GRAPHE DE CONNAISSANCE : LE CAS DE LA CINÉMATHÈQUE

3.1 Introduction

Ce projet pilote, mené par la Cinémathèque québécoise en collaboration avec une équipe de l'École Polytechnique de Montréal, s'inscrit dans le cadre du projet Savoir Commun. Son objectif est de montrer la faisabilité de la publication en données ouvertes liées de certaines informations de la base de données de la Cinémathèque (CineTV). Comme celle-ci contient des informations riches sur plus d'une centaine de milliers d'oeuvres audiovisuelles, concernant, notamment, les personnes et les technologies impliquées dans leur production, la publication d'une partie ou de l'ensemble de son contenu offrirait la possibilité aux citoyens et chercheurs d'avoir libre accès à une partie importante du patrimoine québécois.

L'utilisation de données liées, plus spécifiquement de données encodées en langage RDF (Resource Description Framework), permet d'offrir la possibilité aux utilisateurs de poursuivre leur exploration du monde cinématographique au-delà des connaissances contenues dans les données de la Cinémathèque ; en intégrant des liens vers d'autres ressources en ligne telles que Wikidata, DBpedia ou encore d'autres bases de connaissances sur le cinéma, les outils de recherche utilisés peuvent consolider de manière dynamique les informations de sources diverses liées aux oeuvres considérées.

3.2 Méthodologie de développement

Analyse préliminaire

Plusieurs étapes sont nécessaires pour le développement d'un prototype d'outil de traduction des données en RDF. En premier lieu, une analyse préliminaire des données de la Cinémathèque est cruciale. Cette analyse se fait autant au niveau du contenu que de l'examen des structures et quantités de données de différents types.

Dès le début du projet, plusieurs documents et fichiers de données ont été fournis par la Cinémathèque. Les documents initiaux incluaient les règles de catalogage pour l'entrée de données dans le système actuel, la documentation de correspondance pour un projet parallèle

de changement de base de données et les listes d'autorité utilisées dans CineTV (des listes de valeurs permises pour certains champs, afin d'éviter les inconsistances).

À partir de cette information, un diagramme représentant la structure de CineTV a été reconstruit. Les types de données des champs et leur longueur maximale ont été introduits dans cette représentation afin de pouvoir visualiser facilement quelles parties de CineTV sont peuplées par quel type d'information.

Sélection des données à traduire

Par la suite, la sélection de sous-parties de la base de données à utiliser est nécessaire, puisque l'implémentation de l'entièreté de celle-ci en données liées est un travail d'une envergure dépassant le projet. L'élaboration de « questions de compétences », c'est-à dire de questions pour lesquelles un utilisateur devrait pouvoir obtenir une réponse, permet de cerner quelles parties de la base de données devraient être intégrées de manière prioritaire au modèle développé.

Modélisation d'ontologie

Comme il s'agit de transformer des données sous forme de tables relationnelles vers un graphe de connaissances, un travail de modélisation de la structure souhaitée - l'ontologie - est de mise. Plusieurs ontologies existantes couvrent des domaines apparentés au catalogage d'œuvres cinématographiques, et ont été examinées afin de déterminer si elles peuvent être réutilisées dans le modèle de la Cinémathèque. Lorsque les ontologies existantes retenues sont insuffisantes pour les besoins du modèle, elles doivent être étendues via l'ajout de sous-classes et sous-propriétés afin de répondre aux besoins du projet. Le processus de modélisation utilisé est basé sur la méthodologie SAMOD, présentée à la sous-section 2.2.5. Les décisions prises à ce stade ont d'importantes ramifications sur le type de concepts qui sont mis en valeur dans la structure finale.

Vérification de l'implémentation de l'ontologie

Une fois cette modélisation terminée, l'implémentation de l'ontologie est effectuée. Il est ensuite possible d'écrire les requêtes en langage SPARQL qui permettent de vérifier la cohérence sémantique de cette ontologie. Si l'ontologie est bien modélisée et bien implémentée, l'écriture des requêtes SPARQL visant à extraire des informations devrait être triviale. Ces requêtes permettent également de vérifier, jusqu'à un certain degré, l'intégrité des données,

une fois la traduction effectuée.

Traduction RDF

Après implémentation de l'ontologie, il est nécessaire d'élaborer une méthode de traduction des données à partir des bases relationnelles vers la syntaxe RDF. Un script charge les données initiales, et créé les triplets correspondants du graphe, en concordance avec le modèle déterminé précédemment.

Vérification du graphe final

Finalement, il faut assurer l'intégrité des données et la préservation du sens. Les requêtes SPARQL élaborées après le développement de l'ontologie ainsi que des décomptes des entités principales sont les méthodes principales de vérification.

Ces étapes de développement sont décrites de manière plus détaillée dans les sections suivantes.

3.3 Élaboration des questions de compétence

Lors de discussions avec les équipes de la Cinémathèque, plusieurs sujets d'intérêt, s'apparentant à des scénarios motivants de la méthodologie SAMOD [18], ont été identifiés. Ces scénarios ont été traduits en questions de compétence informelles sous forme de langage naturel. Ces questions de compétence informelles nous ont servis à déterminer quelles données devaient être traduite à partir des données source. De plus, elles nous permettent de déterminer quelles entités doivent être présentes dans le modèle, nous permettant de cerner quelles entités des ontologies d'intérêt existantes intégrer au modèle ontologique développé. Finalement, ces questions serviront également de base à l'implémentation de questions de compétence formelles en langage SPARQL, qui permettent de valider la fonctionnalité de l'ontologie et la validité du contenu de la base de connaissances. Les questions de compétence informelles, regroupées par thèmes, sont les suivantes :

1. Les groupes de personnes ayant souvent travaillé ensemble :
   (a) Quels groupes de femmes, hommes ou autres personnes québécoises ont le plus souvent travaillé ensemble dans différents films ?
   (b) Est-ce qu'une femme, un homme ou autre personne québécoise collabore avec des femmes/hommes/autres personnes québécoises différentes, ou bien travaille-t-elle

plutôt avec les mêmes personnes ?

2. Les fonctions des personnes et leurs évolutions dans le temps :
   (a) Localement : Comment a évolué la carrière d'une personne québécoise spécifique ?
   (b) Globalement : Y a-t-il des « patrons » de carrière (c'est-à-dire des postes cinématographiques typiquement occupés l'un après l'autre) ?
3. Données liées :
   (a) Quelles personnes québécoises se trouvant sur WikiData ont un ID de la Cinémathèque québécoise ?
   (b) Quelles personnes de la base de données de la Cinémathèque se retrouvent sur Wikidata ?

Ces trois classes de questions permettent de démontrer l'utilité des données liées pour identifier de quelle manière des entités sont liées : groupes de personnes, informations temporelles (évolutions de carrières) et informations connexes sur d'autres sites (comparaisons avec Wikidata).

3.4 Choix des informations à conserver

En se basant sur les questions de compétences précédentes, la sous-section de CineTV à intégrer à l'ontologie prototype a été identifiée. Les tables relationnelles nécessaires sont présentées au tableau 3.1.

Nom de la table	Informations utilisées
Filmo	Titre, années de sortie et de production
Filmo_Pays	Pays de production
Filmo_Generique	Fonctions exécutées par les personnes lors de la production
Filmo_Realisation	Réalisateurs de films
Fonction	Termes des fonctions utilisées dans Filmo_Generique
Nom	Prénom, nom, et l'information sur la nationalité québécoise des personnes utilisées dans Filmo_Generique
Pays	Nom des pays utilisés dans Filmo_Pays

Tableau 3.1 Tables relationnelles de CineTV utilisées

À noter que le genre des personnes est nécessaire afin de pouvoir répondre adéquatement aux questions de compétence. Cependant, cette information n'est pas présente dans la base de données CineTV. Il est possible de récupérer les genres des personnes de manière automatisée à partir d'une source externe comme Wikidata, ce que l'on a fait à l'aide d'un script spécialisé.

Une fois les informations nécessaires repérées, l'analyse des ontologies existantes permet d'identifier quels modèles sont suffisants pour les représenter. On peut remarquer que les principaux types d'informations requises sont par rapport aux films, aux gens et aux pays, ainsi que les attributs associés (nom, genre, etc.).

Dans le tableau 3.2 sont regroupées les entités les plus importantes parmi certaines des ontologies intégrées dans notre modèle. Dans la colonne de gauche, on retrouve l'entité du modèle de la Cinémathèque. Dans les colonnes suivantes, il s'agit de la propriété ou de la classe représentant cette même information dans les divers modèles ontologiques retenus. Ces modèles sont décrits plus en détail à la section 3.5. Pour schema.org et DBpedia, une majuscule indique une classe ; une minuscule indique une propriété.

Cinémathèque	**FRBRoo**	**CIDOC**	**schema.org**	**DBpedia**
Filmo	F21_Recording_Work	E73_Information_Object	Movie	Film
Pays	E53_Place	E53_Place	Country	Country
Personne	E21_Person	E21_Person	Person	Person
Nom	E82_Actor_Appellation	E82_Actor_Appellation	name	-
Genre	E55_Type	E55_Type	gender	-
Réalisation	F29_Recording_Event	E7_Activity	-	-
Sortie	E7_Activity	E7_Activity	datePublished	completionDate
Titre	E35_Title	E35_Title	name	originalTitle
Fonction	E55_Type	E55_Type	-	-
NomPays	E44_Place_Appellation	E44_Place_Appellation	name	-
Durée	E52_Time-Span	E52_Time-Span	-	-

Tableau 3.2 Entités retenues dans CineTV et leurs correspondances dans les modèles retenus

FRBRoo réutilise un grand nombre de classes issues de CIDOC-CRM. Par contre, certaines classes sont davantage spécifiées dans FRBRoo, particulièrement celle décrivant les oeuvres enregistrées et l'événement d'enregistrement d'une oeuvre. Les attributs d'appellation de FRBRoo et CIDOC-CRM sont plus riches que ceux de schema.org, qui n'a que la propriété *name*. DBpedia et schema.org n'ont pas d'entités pour représenter les événements (tels que la réalisation), les fonctions qu'occupent les personnes dans la création d'une oeuvre, ou la durée temporelle des événements. Ces différences illustrent la différence de complexité et de spécialisation entre les différents modèles considérés.

3.5 L'ontologie CMTQ

Nous proposons une solution à deux modèles : un modèle simple pour permettre de répondre facilement aux questions de base, et un modèle plus riche pour les questions plus compliquées. Le modèle simple est basé sur le vocabulaire de schema.org, car comme men-

tionné au chapitre 2, ce vocabulaire est parmi les plus couramment utilisés dans le domaine des données liées ouvertes, et sa structure est minimaliste. Ainsi, il sera facilement approchable pour des utilisateurs déjà familiers avec le monde des données liées. Le modèle riche est basé sur FRBRoo, qui est un modèle de l'état de l'art pour les données de catalogues bibliographiques contenant des données telles que celles de la Cinémathèque. Ce modèle plus complexe, également utilisé par d'autres institutions similaires en vocation ailleurs au monde, permet de représenter plus complètement les informations fournies initialement que le modèle simple. La prochaine section présente le modèle riche, et le modèle plus simple est présenté à la section subséquente.

3.5.1 Modèle cinema_FRBRoo

Dans le premier des modèles que nous proposons, cinema_FRBRoo, un maximum d'information est conservé sur chacune des entités. Chaque événement et chaque titre ou nom de personne est une entité à part entière pouvant être décrite davantage. Ainsi, par exemple, il est possible de conserver l'information suivante sur l'événement de réalisation d'un film : l'endroit de cet événement, la durée de cet événement, ainsi que les fonctions exécutées par des personnes dans le cadre de cet événement. Cette manière de représenter les événements temporels permet de conserver plus d'informations à leur sujet que la simple reprise des dates leur étant associées.

Dans la figure 3.1, les entités et propriétés débutant par `cmtq` (Cinémathèque) réfèrent à celles créées pour le prototype, tandis que `frbroo` réfère à FRBRoo et `crm` à CIDOC-CRM. Pour des raisons de lisibilité et d'espace, même s'il ne s'agit pas d'une notation conventionnelle, nous représentons ainsi les classes dans le diagramme : chaque classe contient comme titre le nom de la classe dans l'ontologie cinema_FRBRoo, et la moitié inférieure des carrés représentent les classes de CIDOC-CRM et FRBRoo dont elle dérive. Par exemple, `Filmo` est une sous-classe de `frbroo:F21_Recording_Work`. Les rectangles aux coins arrondis représentent le type de la valeur pour les propriétés qui prennent comme valeur un nombre ou une chaine de caractères. Chaque flèche unidirectionnelle est un « triplet » d'information en RDF, reliant un sujet à un objet via une relation qualifiée (propriété). Chacune de ces flèches est étiquetée avec d'abord la propriété de cinema_FRBRoo, puis la propriété dont elle dérive. Par exemple, la relation entre `Realisation` et `Duree` est `cmtq:moment_realisation`, qui est une sous-propriété de `crm:P4_has_time-span`.

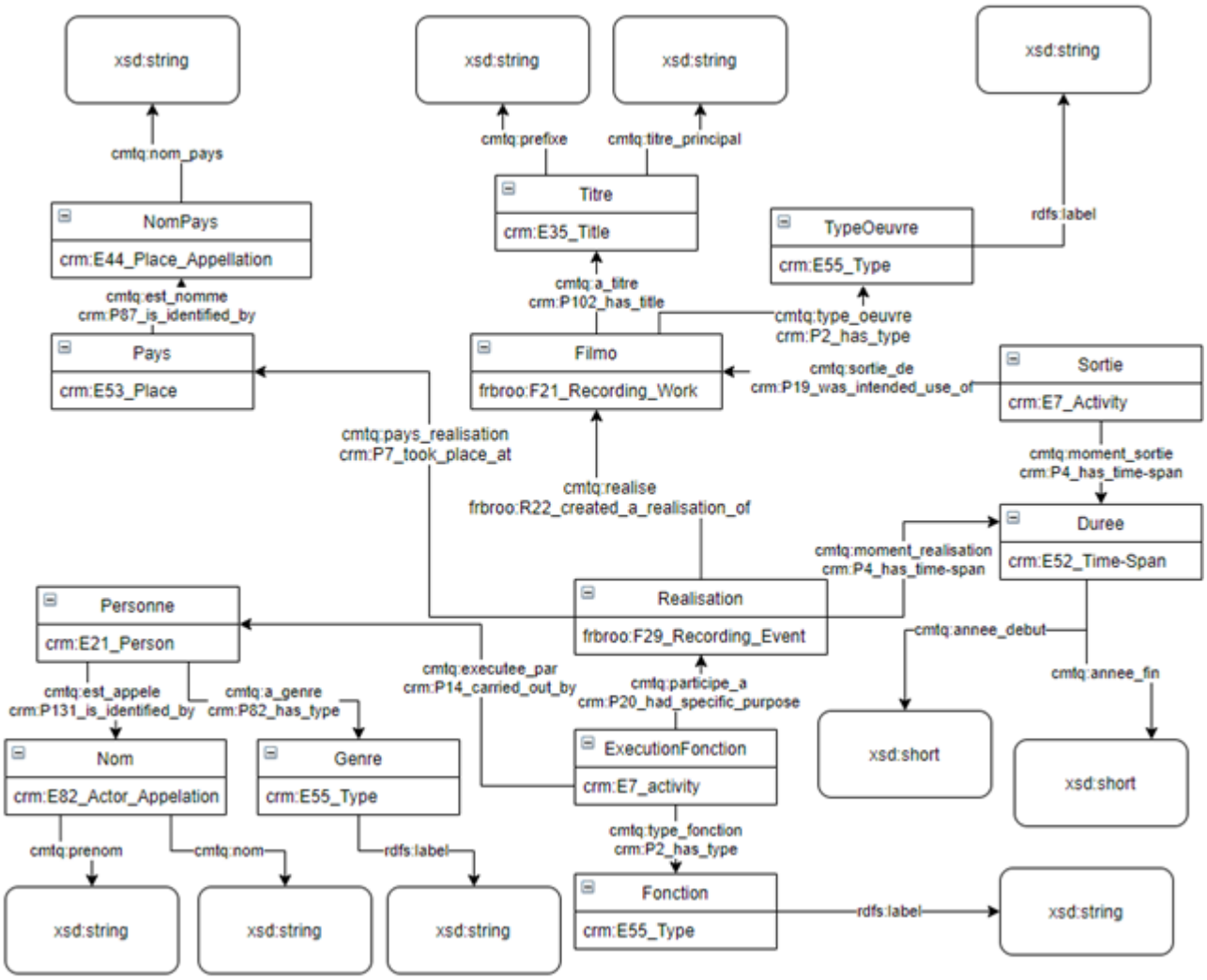

Figure 3.1 Modèle cinema_FRBRoo

Dans l'explication suivante du modèle, les classes du modèle sont représentées avec une majuscule.

Ce modèle est centré sur l'entité de `Realisation`, qui a lieu dans un `Pays` et qui est l'événement englobant toutes les activités nécessaires pour créer un `Filmo`. Ces activités sont représentées par des `ExecutionFonction`, soit des événements représentant le travail d'une `Personne` spécifique, qui exécute une `Fonction` spécifique. Ces `Personnes` sont également liées à un `Genre` et à une `Appellation` (leur nom).

Le `Filmo` est une œuvre conceptuelle produite par cet événement de `Réalisation`. Le `Filmo` est également lié à un événement de `Sortie`, qui correspond au moment où le `Filmo` est présenté au public pour la première fois. De plus, le `Filmo` possède une entité `Titre`.

Pour spécifier la durée des événements, il faut, pour respecter le modèle FRBRoo, déclarer une `Duree`, qui elle aura des moments de début et de fin définis.

Par exemple, voici la représentation RDF du film Incendies et des entités associées du modèle cinema_FRBRoo. Les lignes étant précédées par une double barre oblique sont des commentaires ne faisant pas partie du RDF.

```
// L'entité est un Filmo, et a l'identifiant 78167
cmtq:Filmo78167 a cmtq:Filmo ;
    cmtq:id_cinematheque "78167"^^xsd:short .

// Le film 78167 a un titre. Ce titre a comme valeur « INCENDIES »
cmtq:Filmo78167 cmtq:a_titre cmtq:titre78167 .
cmtq:titre78167 cmtq:titre_principal "INCENDIES"^^xsd:string .

// La sortie 78167 est celle du film 78167 et a un moment de sortie
cmtq:sortie78167 cmtq:sortie_de cmtq:Filmo78167 ;
    cmtq:moment_sortie cmtq:momentSortie78167 .

// Le moment de sortie a commencé en 2010 et termine en 2010
cmtq:momentSortie78167 cmtq:annee_debut "2010"^^xsd:short ;
    cmtq:annee_fin "2010"^^xsd:short .


// Un événement de réalisation est associé au film, et à un moment
cmtq:realisation78167 cmtq:realise cmtq:Filmo78167 ;
    cmtq:moment_realisation cmtq:momentRealisation78167 .
    // Si on avait les dates de réalisation, il y aurait des triplets
    // indiquants ces dates.

// La réalisation a eu lieu dans les pays 105 et 216 (France et Québec)
cmtq:realisation78167 cmtq:pays_realisation cmtq:Pays105 ;
    cmtq:pays_realisation cmtq:Pays216 .

// Une fonction a été exécutée, et son type est Realisation. Elle a
// été exécutée par la personne 37191, et l'exécution de cette
// fonction est un événement faisant partie de l'événement de réalisation
```

```
cmtq:generique72610 a cmtq:ExecutionFonction ;
    cmtq:type_fonction cmtq:Realisation ;
    cmtq:executee_par cmtq:Personne37191 ;
    cmtq:participe_a cmtq:realisation78167 .


// La personne 37191 a l'identifiant Cinémathèque 37191, et a
// comme étiquette « Denis Villeneuve »
cmtq:Personne37191 a cmtq:Personne ;
    cmtq:id_cinematheque "37191"^^xsd:short ;
    rdfs:label "Denis Villeneuve"^^xsd:string .

// La personne 37191 a une appellation. Cette appellation est composée du
// prénom « Denis » et du nom « Villeneuve ». La personne est de genre
// masculin.
cmtq:Personne37191 cmtq:est_appele cmtq:appellation37191 ;
    cmtq:a_genre cmtq:Masculin .
cmtq:appellation37191 cmtq:prenom "Denis"^^xsd:string ;
    cmtq:nom "Villeneuve"^^xsd:string .


// Le pays 105 a une appellation. Cette appellation a comme valeur « France »
cmtq:Pays105 a cmtq:Pays ;
    cmtq:est_nomme cmtq:appellation_pays105 .
cmtq:appellation_pays105 cmtq:nom_pays "France"^^xsd:string .
```

3.5.2 Modèle simplifié

Le modèle simplifié, qui sera utilisé en conjonction avec le modèle présenté à la section précédente, consiste en la modélisation minimale nécessaire pour être en mesure de répondre aux questions de compétence. Ainsi, par exemple, au lieu de représenter la sortie d'un film par une entité à part entière, on ne conserve que l'année de sortie en tant qu'attribut de la classe `Filmo`. Ceci permet d'écrire plus rapidement des requêtes, qui auront une meilleure performance. Par contre, l'expressivité potentielle de ce modèle est réduite. Entre autres, au lieu d'avoir une entité « Titre » qui peut contenir l'article défini et la partie principale du titre séparément (pour des fins de classification alphabétique), le titre est exprimé comme attribut simple d'un film.

Les vocabulaires DBpedia, schema.org, RDA et Dublin Core, étant répandus, sont intégrés au modèle simplifié afin de faciliter l'exploitation de l'ontologie par des experts familiers avec ceux-ci, et l'interopérabilité avec des ontologies les utilisant. DBpedia est représenté par le préfixe *dbo*, schema.org par *schema*, RDA par *rdac* et Dublin Core par *dcterms*.

Là où des propriétés des modèles simples et cinema_FRBRoo ont la même sémantique (par exemple, « a pour titre »), sans être implémentées de la même manière (avec un littéral dans le cas du modèle simple, et une entité dans cinema_FRBRoo), le suffixe `_lit` est ajouté à la propriété du modèle simple (lit pour « littéral ») pour les distinguer.

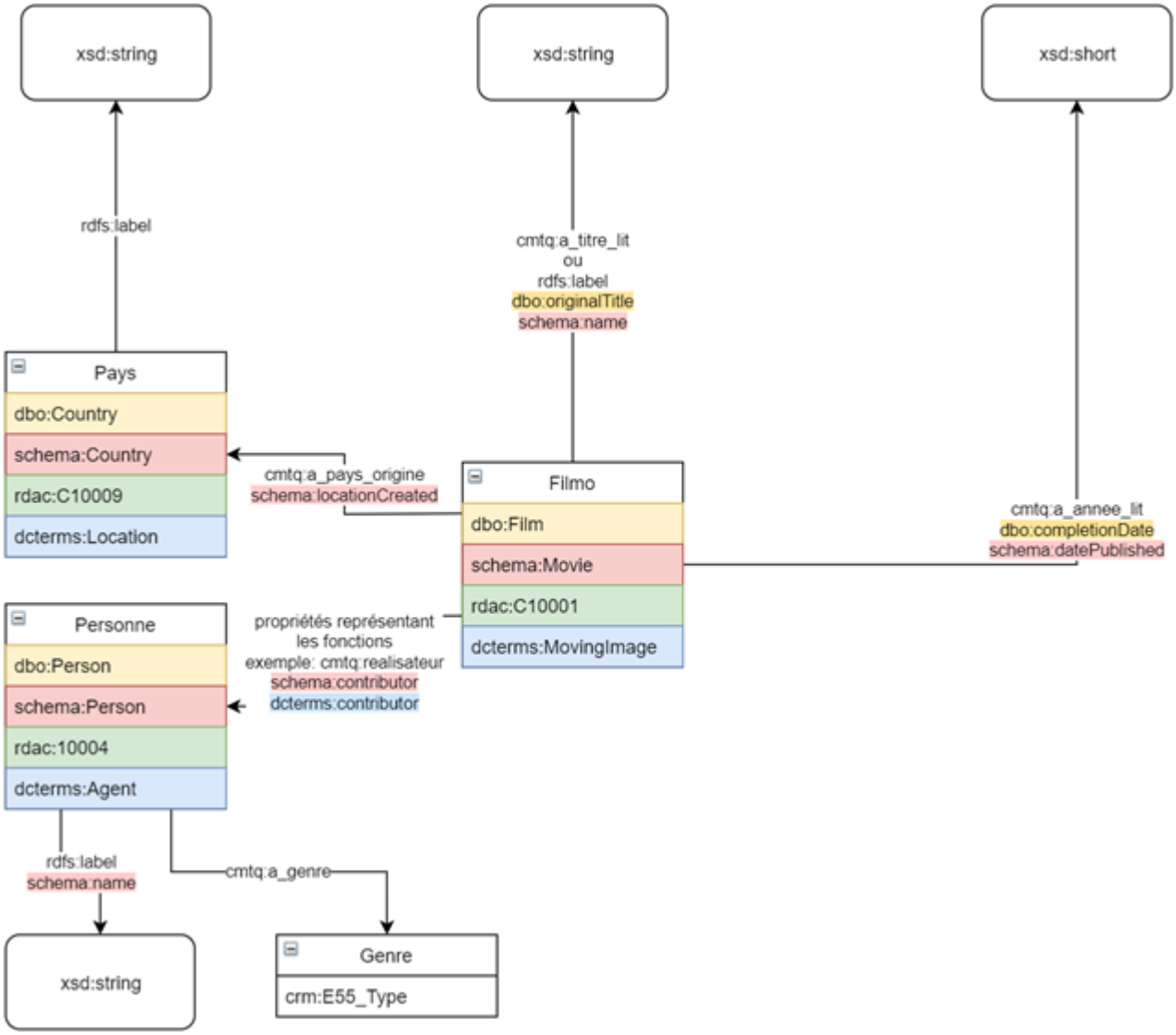

Figure 3.2 Modèle simplifié

Dans ce modèle, il n'y a que trois entités principales : `Filmo`, `Pays` et `Personne`. Le reste de l'information est encodé avec des propriétés littérales. Par exemple, en contraste avec le modèle cinema_FRBRoo, la date de sortie d'un `Filmo` est simplement un attribut de ce dernier, au lieu d'être un attribut de la `Duree` de l'événement de `Sortie` du `Filmo`. Ainsi, il est possible d'écrire des requêtes plus simples pour l'exploitation des données en utilisant ce modèle, ce qui permet des gains en performance. Cependant, ceci vient aux dépens du pouvoir expressif du modèle quant aux entités retirées.

```
// Le film 78167 est un Filmo. Il a comme titre « INCENDIES » et
```

```
// est paru en 2010.
cmtq:Filmo78167 a cmtq:Filmo ;
    cmtq:a_titre_lit "INCENDIES"^^xsd:string ;
    cmtq:a_annee_lit "2010"^^xsd:short ;

    // Les pays d'origine du film sont la France et le Québec.
    cmtq:a_pays_origine cmtq:Pays105 ;
    cmtq:a_pays_origine cmtq:Pays216 ;


// La personne 37191 a réalisé le film 78167
cmtq:Filmo78167 cmtq:realisateur cmtq:Personne37191 .


// La personne 37191 est une personne. Elle s'appelle Denis Villeneuve et
// est de genre masculin.
cmtq:Personne37191 a cmtq:Personne ;
    rdfs:label "Denis Villeneuve"^^xsd:string ;
    cmtq:a_genre cmtq:Masculin .


// Le pays 105 est nommé « France »
cmtq:Pays105 rdfs:label "France"^^xsd:string .
```

3.5.3 Juxtaposition des modèles

La figure 3.3 illustre le lien entre les deux modèles précédents. Les entités en rouge représentent les classes conservées dans la version simplifiée ; les flèches rouges représentent les relations simplifiées – on peut remarquer que, souvent, deux propriétés et une classe intermédiaire sont sautées, parfois plus. De plus, les champs de données verts, jaunes et bleus sont concaténés : préfixe avec titre, prénom avec nom, et une année, soit l'année de sortie, si elle est présente dans CineTV ; sinon, les années de fin ou de début de production. Ainsi, au lieu d'avoir deux attributs pour représenter le titre, il n'y en a qu'un seul, et il n'y a qu'un attribut année au lieu de trois, simplifiant l'exploitation.

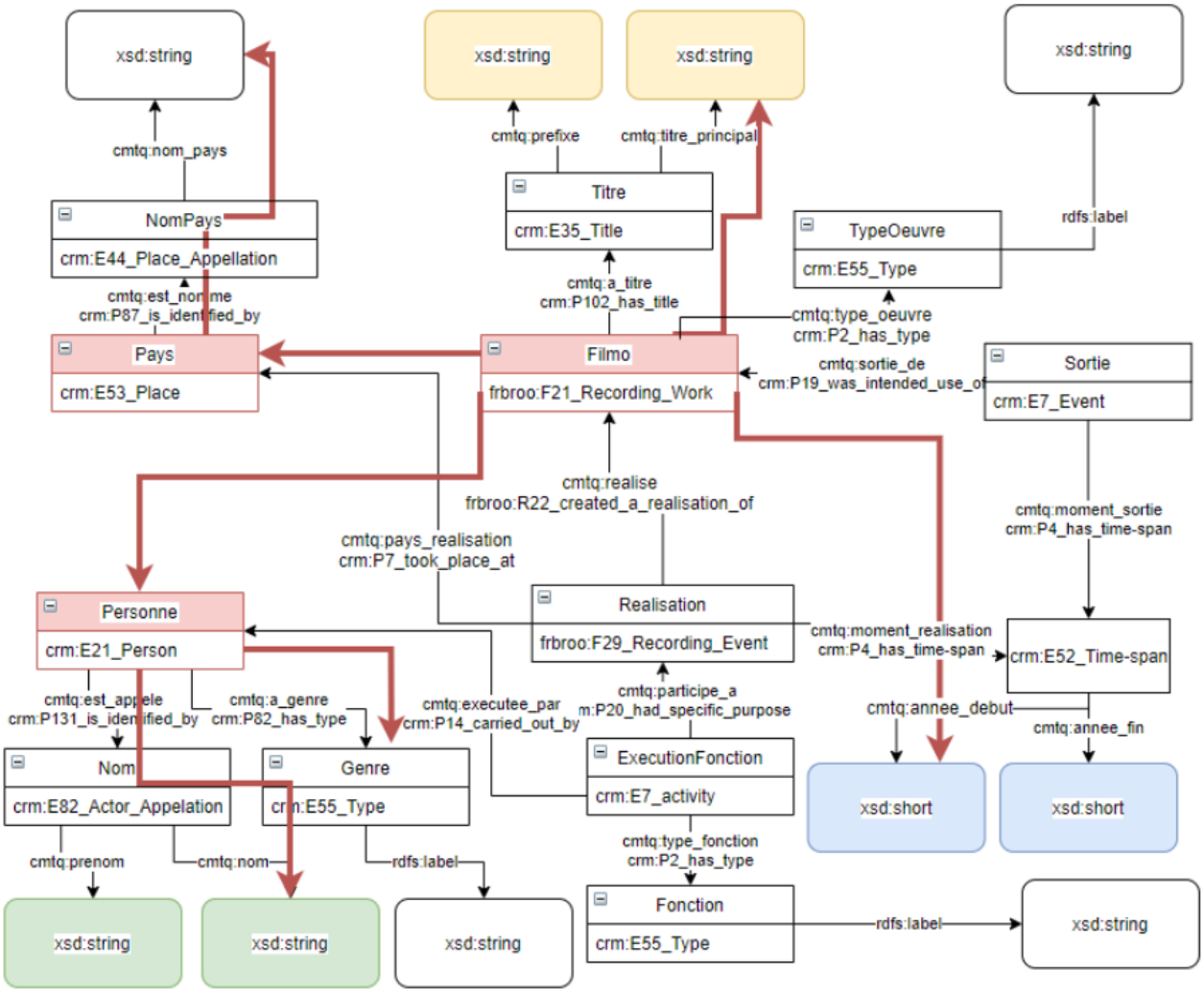

Figure 3.3 Juxtaposition des modèles cinema_FRBRoo et simplifié

Ainsi, lors de l'exploitation du modèle cinema_FRBRoo, pour connaître l'année de sortie d'un film, il faut se poser la question suivante : quelle est l'année associée au laps de temps de l'événement de sortie associé au film ? En utilisant les raccourcis du modèle simple, cette question devient tout simplement : quelle est l'année de sortie du film ?

3.6 Conversion des données

Une fois la portion de données originales à conserver choisie et l'ontologie élue et implémentée, la conversion des données du format original relationnel au format final RDF est effectuée. Pour ce faire, un script Python a été élaboré qui prend en entrée les fichiers relationnels originaux fournis, au format de feuilles de calcul .xslx. Le script effectue la véri-

fication du formatage et contenu des champs, puis écrit dans le fichier destination les triplets RDF nécessaires afin de représenter les entités dans le graphe de connaissances.

3.7 Enrichissement

Certaines informations absentes de fichiers sources sont nécessaires pour pouvoir répondre aux questions de compétence dérivées des demandes des acteurs du domaine. Notamment, les genre des personnes impliquées est d'une haute importance. Grâce à un travail antérieur de versement des identifiants de la Cinémathèque dans la base de connaissances Wikidata, extraire de l'information supplémentaire de cette source de manière automatique est possible. Ainsi, un script supplémentaire est en mesure d'interroger Wikidata pour y récupérer le genre des personnes, lorsque cette information s'y retrouve.

3.7.1 Désambiguïsation

Certaines des entités décrites dans les données originales sont des entités déjà décrites de manière extensive par d'autres bases de connaissances, notamment les pays. Selon les principes des données liées ouvertes, il est donc préférable d'intégrer à notre graphe de connaissances des références non-ambiguës vers ces ressources existantes, plutôt que de répéter l'information disponible. Un script effectuant des requêtes à l'API des lieux de Google nous permet ainsi de récupérer les identifiants uniques assignés à chaque endroit mentionné dans les données originales et les intégrer au graphe. Ainsi, une requête peut retourner non seulement le nom du lieu associé à un événement, mais également un lien direct vers une source riche en informations plus poussées sur le lieu en question.

De plus, l'information de notre graphe est enrichie par le fait que la source externe, Google Places dans le cas présent, organise les lieux selon une taxonomie hiérarchique. En intégrant cette taxonomie à notre graphe, la porte est ouverte à de nouvelles requêtes plus riches. Par exemple, même si le graphe ne mentionne pas de manière explicite la ville de naissance d'une personne, une requête sur les personnes nées dans le pays où se situe la ville retournera les informations par rapport à cette personne, étant donné que la taxonomie spécifie que la ville se situe dans une province qui elle-même se situe dans le pays en question.

3.8 Évaluation du graphe de connaissances

Une revue des méthodes d'évaluation d'ontologies effectuée par Raad et Cruz en 2015 [28] propose des méthodes pour vérifier si l'ontologie développée présente certaines caractéris-

tiques nécessaires. La comparaison à une ontologie de référence existante et vérifiée (*gold standard*) ou encore au contenu d'un corpus de référence est une de ces méthodes, permettant de mesurer la précision, la complétion et la concision. Deux autres méthodes d'évaluation sont l'évaluation experte basée sur des critères de cohérence logique et l'évaluation de la capacité de l'ontologie d'accomplir des tâches. Ces deux méthodes peuvent servir à vérifier l'adaptabilité, la clarté, l'efficacité de calcul et la consistance de l'ontologie modélisée.

Les méthodes d'évaluation formelles présentées par cette revue étant trop coûteuses en temps ou en ressources pour l'envergure du projet, elle n'auraient pu être employées. Cependant, des versions moins rigoureuses de ces méthodes ont été employées afin de vérifier ces mêmes critères dans la mesure du possible. La comparaison à un corpus de référence a pu être effectuée par le décompte d'entités identiques entre la base de données originales, CineTV, et le graphe de connaissances final. La vérification par les tâches est analogue à la réponse à des questions de compétence, type de validation également utilisé par la méthodologie SAMOD. Finalement, une vérification par des acteurs du milieu a été effectuée grâce au développement par un autre étudiant d'un prototype d'application de visualisation des données permettant aux acteurs d'explorer le graphe de connaissances via une interface graphique.

3.8.1 Questions de compétence

Les questions de compétence sont développées à partir des scénarios motivants des utilisateurs. Avant le développement d'une partie du modèle, les questions de compétence auquel cette partie doit pouvoir répondre sont posées en langue naturelle, comme présenté à la section 3.3. Les questions de compétence permettent d'évaluer l'aptitude du modèle à répondre à des tâches utilisateur du monde réel. Les questions de compétence sont traduites en requêtes SPARQL, pour vérifier si le modèle est en mesure de répondre aux scénarios motivants.

Par exemple, la question de compétence 2 s'intéresse aux fonctions que les personnes ont occupées et leur évolutions dans le temps. Une des questions issues de ce scénario est la suivante : "Y a-t-il des patrons de carrière (c'est à dire des postes cinématographiques typiquement occupés l'un après l'autre) ?" La traduction de cette requête en SPARQL est présenté à la figure 3.4 .

```
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX cmtq: <https://data.cinematheque.qc.ca/data#>


select ?fonction (count(distinct ?participant) as ?instances)
 where {
    ?participant a cmtq:Personne .

    ?filmo1 a cmtq:Filmo .
    ?filmo1 cmtq:a_titre_lit ?titre1 .
    ?filmo1 ?fct1 ?participant .
    ?filmo1 cmtq:a_annee_lit ?as1 .

    ?filmo2 a cmtq:Filmo .
    ?filmo2 cmtq:a_titre_lit ?titre2 .
    ?filmo2 ?fct2 ?participant .
    ?filmo2 cmtq:a_annee_lit ?as2 .

    ?fct1 rdfs:label ?fonction1 .
    ?fct2 rdfs:label ?fonction2 .
    FILTER(?filmo1 != ?filmo2)
    FILTER(?as1 < ?as2)
    FILTER(?fct1 != ?fct2)
    BIND(concat(?fonction1, ", ", ?fonction2) as ?fonction)
} group by ?fonction order by desc(count(distinct ?participant))
 LIMIT 10000
```

Figure 3.4 Requête SPARQL pour répondre à la question de compétence 2b

Cette requête, lorsqu'elle interroge la base de connaissances finale, retourne les paires de fonctions qui se suivent le plus souvent dans les carrières d'une personne ainsi que le nombre de fois que ces fonctions ont été occupées l'une après l'autre. Les cinq premiers résultats retournés sur la base de connaissances sont présentés à la figure 3.5.

	fonction	instances
1	Scénario, Réalisation	"2882"^^xsd:integer
2	Réalisation, Scénario	"2859"^^xsd:integer
3	Réalisation, Producteur	"1766"^^xsd:integer
4	Producteur, Réalisation	"1597"^^xsd:integer
5	Interprétation, Réalisation	"1428"^^xsd:integer

Figure 3.5 Cinq premiers résultats retournés par la requête 3.4

L'implémentation de toutes les questions de compétence ainsi que les premiers résultats retournés sont présentés à l'annexe A.

Les requêtes des figures 3.6 et 3.8 récupèrent toutes les informations brutes nécessaires pour répondre à ces questions, et sont utilisées par le protoype d'application web pour récupérer auprès de la base de connaissances les données nécessaires à l'affichage. Une de ces requêtes est effectuée sur le modèle cinema_FRBRoo, et l'autre sur le modèle simplifié, afin d'illustrer la différence de complexité entre les deux.

La requête de la figure 3.6 permet de récupérer toutes les informations de base nécessaires pour répondre aux questions de compétences énoncées à la section 3.3. Elle récupère, dans l'ordre, les informations suivantes :

1. Tout Filmo de la base de connaissances, ainsi que son titre.
2. Si la partie principale d'un titre est précédé par un article, l'article. Par exemple « La » dans le titre « La Belle et la Bête ».
3. La partie principale du titre. Par exemple, « Belle et la Bête ».
4. L'événement de sortie du Filmo.
5. Le moment temporel concernant cet événement.
6. L'année du moment de la sortie, si elle est présente.
7. L'événement de réalisation du Filmo.
8. L'identifiant unique (IRI, *Internationalized Resource Identifier*) du pays où a eu lieu cet événement.
9. Le moment temporel de cet événement.
10. Si disponibles, les années de début et de fin de ce moment.

11. L'appellation du pays de réalisation du Filmo, ainsi que la valeur textuelle de celui-ci.
12. La fonction de chaque événement d'exécution de fonction.
13. La participation de cet événement d'exécution à l'événement de réalisation.
14. L'identifiant de la personne ayant exécuté la fonction.
15. La valeur textuelle du nom de la fonction. (Par exemple, « Maquillage »).
16. L'appellation de la personne ayant exécuté la fonction.
17. La valeur textuelle du prénom de cette appellation.
18. La valeur textuelle du nom de famille de cette appellation.
19. Le genre de la personne.

Les informations conservées à la fin sont les titres, les noms des personnes et pays, les genres des personnes et les noms des fonctions et les années.

Dans la requête de la figure 3.6, les propriétés et classes avec le préfixe `cmtq` représentent des sous-propriétés et sous-classes de CIDOC-CRM et FRBRoo, mis à part les propriétés de données qui ne figurent pas dans ces ontologies. Ces propriétés et classes pourraient être remplacées par leur superclasses (voir la figure 3.3) ; le modèle est, par conséquent, compatible avec ces ontologies.

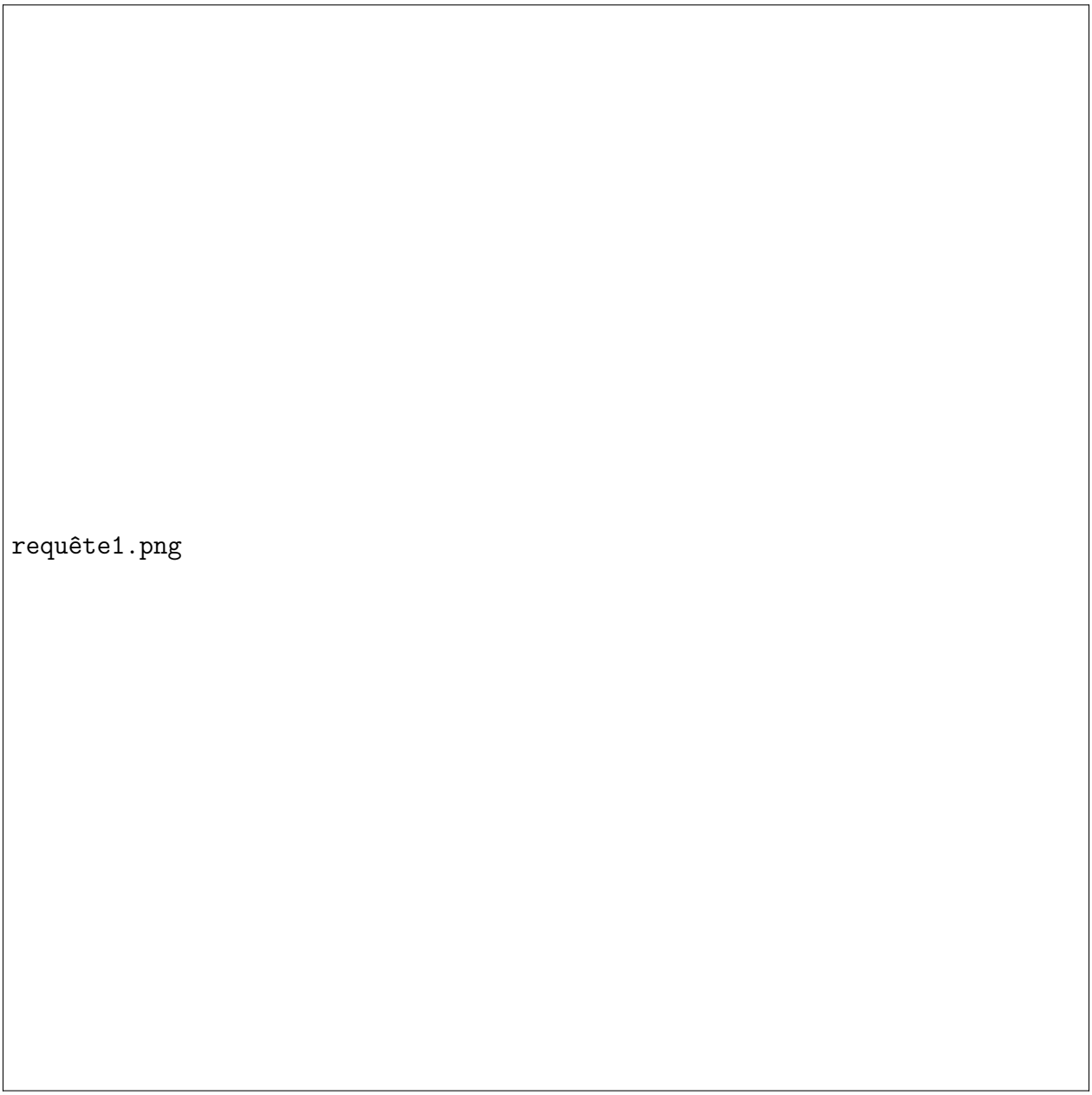

Figure 3.6 Requête sur le modèle cinema_FRBRoo

	prefixe	titre_principal	AnneeSortie	Nom	AnneeDebRealisation	AnneeFinRealisation	Fonction	Pays	Genre
1		LOVE IN THE AFTERNOON	"1957"^^xsd:short	A.L. Diamond	"1957"^^xsd:short		Scénario	États-Unis	cmtq:Feminin
2	THE	TARNISHED ANGELS	"1957"^^xsd:short	Agostino Carlesso	"1957"^^xsd:short		Realisation	États-Unis	cmtq:Masculin
3	THE	TARNISHED ANGELS	"1957"^^xsd:short	Agostino Carlesso	"1957"^^xsd:short		Interprétation	États-Unis	cmtq:Masculin
4	LE	GRAND BLEU	"1988"^^xsd:short	Aki Kaurismäki	"1988"^^xsd:short		Realisation	États-Unis	cmtq:Masculin
5	LE	GRAND BLEU	"1988"^^xsd:short	Aki Kaurismäki	"1988"^^xsd:short		Direction artistique	États-Unis	cmtq:Masculin

Figure 3.7 Résultats partiels de la requête de la figure 3.6

La requête de la figure 3.8 récupère la même information que la requête précédente (fig. 3.6), mais sur le modèle simplifié. En revanche, on constate que le nombre de lignes est réduit de plus de moitié, passant de 22 à 9 triplets. Cette simplification permet non seulement une exploitation intuitive, mais réduit le temps de traitement nécessaire, chaque ligne étant une nouvelle instruction à exécuter pour le moteur de requêtes.

requête2.png

Figure 3.8 Requête sur le modèle simplifié

Bien que le modèle simplifié suffise pour récupérer l'information brute pour cette application, il n'est pas en mesure de, par exemple, stocker séparément le préfixe et la portion principale du titre d'un film, de représenter les prénoms et noms d'un acteur comme étant les attributs de son appellation ou encore de distinguer entre les événements de sortie et de réalisation d'un film, et d'associer l'implication des personnes à ces événements précis. Ainsi, la juxtaposition des deux modèles permet à la fois d'avoir une certaine convivialité pour les questions simples et récurrentes, mais d'avoir la richesse de représenter les complexités des événements entourant la réalisation et la sortie d'un film.

3.8.2 Vérification de l'intégrité des données

Afin d'assurer la complétude et l'exactitude des données suite à traduction des données des fichiers source vers le graphe final, le décompte des entités originales de CineTV a été effectué, puis été comparé avec le décompte des entités du graphe de connaissances final. Les décomptes des ces entités de la base de données CineTV et de la base de connaissances finale sont présentés au tableau 3.3.

Entité	Filmo	Sujet	Fonction	Personne	Pays
CineTV	104 392	24 072	47	142 767	379
cinema_FRBRoo	104 392	24 072	48	142 767	379

Tableau 3.3 Décompte des entités dans les fichiers fournis et le graphe final

Ces résultats indiquent que l'ensemble des entités présentes dans les données source a correctement été traduite vers le graphe final. La différence au niveau du décompte de fonctions s'explique par l'ajoute de la fonction *Réalisation*, absent dans CineTV, au modèle cinema_FRBRoo.

3.8.3 Prototype d'application

Un prototype d'application web, développé par François Lévesque sous la direction du professeur Thomas Hurtut, consiste en une interface graphique permettant d'explorer le graphe de connaissances, et inclut notamment des représentations visuelles des questions de compétence. Les acteurs du milieu, représentés par le personnel de la Cinémathèque et de la BAnQ, ont pu valider l'utilité et la pertinence du graphe de connaissances à travers l'utilisation de celle-ci. L'application utilise des techniques de visualisation de l'état de l'art. La figure 3.9 consiste en une capture d'écran d'une portion du prototype.

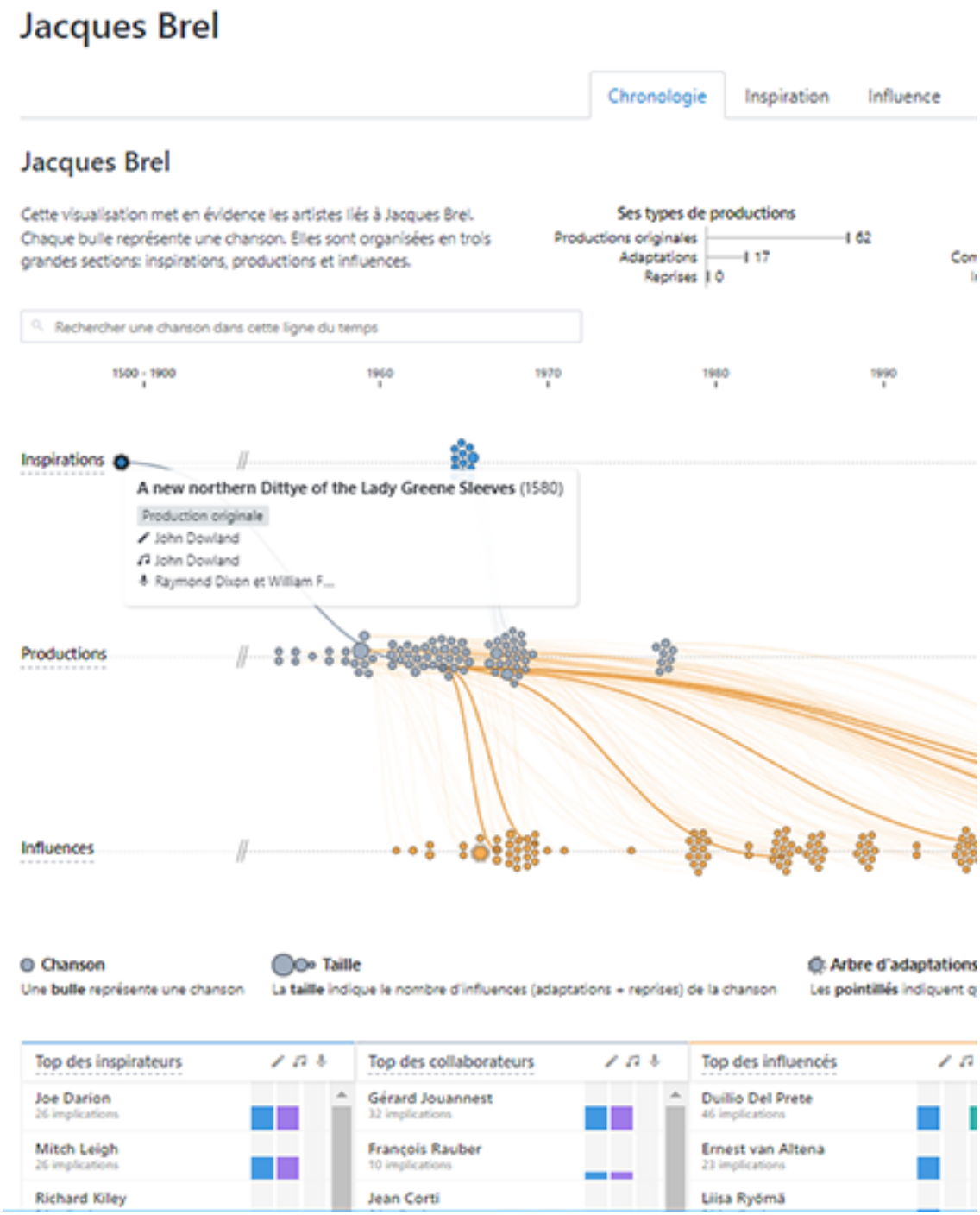

Figure 3.9 Capture d'écran d'une portion du prototype d'application

La réponse aux questions de compétence et l'utilisation de la base de connaissances comme source de données pour le prototype démontrent donc que l'ontologie élaborée et les connaissances représentées dans le graphe ont les caractéristiques nécessaires pour mener à bien les

objectifs du projet.

3.9 Conclusion

Un travail de modélisation, traduction et vérification a permis de réaliser une ontologie des connaissances des données du milieu du cinéma. Le modèle développé est cohérent et compatible avec d'autres ontologies répandues. La représentation utilisée par CIDOC-CRM pour les événements permet de représenter précisément les complexités de la réalisation de films. Notre modèle permet de spécifier de manière non-ambiguë où et quand la réalisation a eu lieu, quelles personnes ont exécuté quelles fonctions, ainsi que les subtilités dans les appellations des oeuvres et des personnes. Le modèle simplifié basé sur le schema.org permet d'accéder aux informations de base plus aisément, mais ne conserve pas la richesse du modèle complet. De plus, ces deux modèles implémentent des vocabulaires et des modèles ontologiques dont l'utilisation est déjà répandue dans le monde des données liées ouvertes, facilitant l'exploitation de la base de connaissances pour les utilisateurs du milieu. Donc, le modèle à la base du graphe de connaissances remplit les objectifs initiaux, c'est-à-dire la préservation de l'information source, grâce à la richesse du modèle CIDOC-CRM et la facilité d'exploitation, grâce au modèle simplifié et à l'utilisation de modèles connus.

Les données présentes dans le graphe de connaissances sont complètes et cohérentes avec les données sources, comme vérifié par les réponses aux questions de compétence et les requêtes de décompte d'entités et d'attributs. Des sources externes ont été exploitées pour enrichir les données originales ; les lieux ont été désambiguïsés à l'aide de l'API Google Places, et les genres des personnes ont pu être récupérés de Wikidata.

La combinaison du contenu et de la structure du graphe permet ainsi aux utilisateurs d'effectuer les tâches souhaitées en termes d'exploration des données de ce milieu.

Par contre, les données provenant d'une seule source homogène, certains défis qu'il serait possible de rencontrer dans ce type de projet n'ont pas dû être relevés. De prime abord, étant donné que l'ensemble des données finales, mis à part celles obtenues par le processus de désambiguïsation des lieux, proviennent d'une unique source de données, elles sont sensiblement formatées de manière consistante, avec peu de variabilité dans la structure du contenu, ce qui minimise le travail de nettoyage nécessaire. Ensuite, comme les tables relationnelles fournies utilisent des clés de jointure pour les relations entre entités d'un même type et entités de types différents, aucun travail supplémentaire n'est nécessaire pour assurer qu'on ne se retrouve pas avec plusieurs copies disjointes d'une même entité, puisqu'elles possèdent toutes des identifiants locaux uniques. Ainsi, il n'y a pas la nécessité d'effectuer un travail

d'alignement d'entités au sein de ce graphe. De plus, un versement préalable de certains identifiants dans la base de connaissances Wikidata permet d'aller y récupérer des informations supplémentaires nécessaires aux tâches utilisateur. Finalement, puisqu'il n'y a qu'une seule source de données principale, la conception d'un modèle en mesure de conserver de manière complète l'information s'y trouvant est facilitée. Dans le cas de plusieurs sources initiales, il est plus complexe d'établir une structure qui pourra agglomérer plusieurs visions différentes d'un aspect du monde réel.

Le projet décrit au chapitre 4, qui concerne les données du monde de la littérature québécoise, doit composer avec ces défis et y pallier.

CHAPITRE 4 GÉNÉRATION D'UN GRAPHE DE CONNAISSANCE UNIFIANT DES DONNÉES DE SOURCES HÉTÉROGÈNES : LE CAS DU MCCQ

Dans le chapitre précédent, nous traitions de la modélisation d'une ontologie et de la génération d'un graphe de connaissances pour le domaine du cinéma québécois à partir d'une base de données relationnelle. Ce chapitre traite du processus d'extraction d'une base de connaissances pour le monde de la littérature québécoise.

Les données sources sont composées d'ensembles distincts et hétérogènes provenant de divers acteurs du milieu, ce qui ajoute plusieurs défis. L'ontologie modélisée devrait être compatible avec des modèles répandus, doit être facilement interrogeable et fonctionnelle, et devrait pouvoir conserver et consolider un maximum d'informations de chaque source.

Le processus entier de modélisation et de génération du graphe de connaissances est décrit dans le présent chapitre. La section 4.1 énonce le problème et établit les objectifs. La section 4.2 présente la méthode proposée pour les atteindre, et dresse un portrait des données disponibles. La section 4.3 décrit l'ontologie finale, tandis que la section 4.4 explique les défis et les solutions trouvées en lien avec le processus de traduction des données sources en graphe de connaissance RDF. Finalement, la section 4.5 présente les différentes évaluations de la base de connaissances finale.

4.1 Problématique

Le projet, une collaboration du ministère de la Culture et de Communications du Québec (MCC) et Polytechnique, vise à générer une base de connaissances du monde du livre québécois. Pour ce faire, divers acteurs du milieu ont été réunis, mettant à disposition leurs métadonnées du domaine. Ainsi, des données de Bibliothèque et Archives nationales du Québec (BAnQ), des éditions Hurtubise, de Messageries ADP et de l'Infocentre littéraire des écrivains (L'Ile) ont été fournies. Les ensembles sont encodés en divers formats, et sont de tailles variées.

L'objectif est de réunir ces données dans un seul graphe sous format de données ouvertes et liées. Cet objectif nécessite l'élaboration d'un modèle ontologique pouvant représenter des données bibliographiques de manière à répondre aux cas d'utilisation des acteurs du milieu.

4.1.1 Cas d'utilisation

Au début du projet, les partenaires ont fourni une liste de demandes quant au contenu de la base de connaissances. Parmi les éléments les plus importants figurent :

— les métadonnées des oeuvres ainsi que leurs auteurs, éditeurs et lieux et dates de publication
— les informations quant aux écrivain.e.s telles que leur dates de naissance et leurs genres
— les référents géographiques et temporels, ainsi que les références littéraires/musicales/culturelles présentes dans les oeuvres
— les caractéristiques des éditions d'oeuvres, telles que le nombre de pages et le prix
— les prix littéraires obtenus

La liste complète des demandes formulées par les partenaires se trouve à l'annexe B.

Questions de compétence À partir de ces requis pour le type d'information, des *questions de compétence* ont été établies. Ces questions de compétence sont des questions en langue naturelle ayant trait au domaine du livre québécois. L'objectif est donc de pouvoir répondre à ces questions à partir des connaissances contenues dans le graphe. Les questions de compétence sont les suivantes :

1. Où est né cet auteur ?
2. Où vit cet auteur ?
3. Quand est né cet auteur ?
4. Quand cet auteur est-il mort ?
5. Qui sont les auteurs masculins et féminins ?
6. Quels sont les livres importants écrits par un auteur ?
7. Y a-t-il des références géographiques dans ce livre ?
8. Y a-t-il des références temporelles (année, événements historiques, etc.) dans ce livre ?
9. Y a-t-il des références littéraires / musicales / culturelles dans ce livre ?
10. Quelle est la thématique principale de ce livre ?
11. Ce livre a-t-il été adapté ?
12. Cet auteur a-t-il gagné un prix littéraire, et pour quel livre ?
13. Quels sont les auteurs les plus lus par région ?
14. Dans quels pays ce livre est-il vendu ?
15. Qui est l'illustrateur de ce livre ?

16. Quel est le format de publication de ce livre (gros caractères, format poche) ?
17. Combien de pages ce livre comprend-il ?
18. Ce livre a-t-il des illustrations ?
19. Combien coûte ce livre ?
20. Ce livre a-t-il été auto-publié ou est-il publié commercialement ?
21. Cet auteur a-t-il publié dans un journal ?
22. Quelle est la date de publication de ce livre ?
23. Quel est le lieu de publication de ce livre ?
24. Quelle est la spécialité littéraire de cet auteur ?
25. Combien d'exemplaires de ce livre ont été vendus en ligne à l'étranger ?
26. Ce livre existe-t-il en format numérique ?
27. Quels sont les livres traitant des pratiques numériques ?
28. Cet auteur a-t-il participé à des entrevues / émissions ?
29. Quels sont les livres sur le sujet (exemple : changement climatique) ?

Ces questions de compétence seront traduites en SPARQL (*SPARQL Protocol And RDF Query Language*) une fois l'ontologie définie, afin d'être exécutables sur le graphe de connaissances final. Certains des cas d'utilisation identifiés requerront des données additionnelles pour supplémenter celles fournies par les partenaires.

4.2 Méthodologie

Le processus suivant est proposé pour l'élaboration de l'ontologie et la construction du graphe de connaissances :

— Acquisition des données sources : La première étape consiste à recueillir les ensembles de données des partenaires, sous la forme de fichiers de données ou via une API web.
— Analyse des données sources (4.2.1) : Les modèles de données sous-jacents sont établis par le biais de la consultation de la documentation (s'il y a lieu) et une exploration des ensembles par l'entremise d'outils informatiques analytiques. On analyse la quantité, la structure et la qualité des données sources.
— Définition de l'ontologie (4.3) : Une ontologie est modélisée en tenant compte de plusieurs facteurs, dont les cas d'utilisation, le contenu et la structure des données sources et la compatibilité avec d'autres ontologies existantes du milieu culturel. Les questions de compétence sont traduites en SPARQL, selon la structure du modèle.

— Traduction des données (4.4) : Les données sont extraites de leurs entrepôts sources, nettoyées et formatées, puis sérialisées en RDF. En plus de celles fournies par les partenaires, des données complémentaires provenant de sources ouvertes disponibles sur le web sont retrouvées. Les entités des différentes sources de données sont identifiées et alignées entre elles.
— Évaluation du graphe de connaissances (4.5) : On vérifie que le graphe final correspond aux requis et objectifs initiaux. D'abord, la base de connaissances est chargée sur un serveur avec un module de requêtes SPARQL, puis interrogée avec les questions de compétence. Ensuite, la base de connaissances est évaluée de manière qualitative par des membres des groupes partenaires par l'entremise d'un prototype d'application web permettant d'interagir de manière plus conviviale avec cette dernière.

Ces étapes regroupent chacune plusieurs tâches, méthodes et sous-objectifs, décrits dans les sections correspondantes du présent chapitre.

4.2.1 Présentation des données initiales

Les partenaires du projet ont fourni quatre ensembles de données, soit les données de l'Ile, des Éditions Hurtubise, des Messageries ADP et du dépôt légal de BAnQ. L'objectif de l'analyse de ces ensembles est d'y identifier les entités, leur nombre, propriétés et relations, et de déterminer quels sous-ensembles sont nécessaires afin de répondre aux questions des utilisateurs. Les caractéristiques variées des différents jeux de données sont décrites dans cette section.

L'Ile L'Ile (Infocentre littéraire des écrivains québécois), un ensemble de données provenant du site litterature.org, contient des informations sur plus de 27 000 livres et 1896 auteurs. Le format des données fourni est un tableur sous format csv (*Comma Separated Values*). Chaque entrée concerne un livre, et contient les champs suivants :

— **ID**, contenant un identifiant unique spécifique à l'Île ;
— Le **titre** du livre ;
— La **date de publication**
  Note : il y a parfois deux dates de publication. Il semblerait que la première corresponde à la "vraie" date, et la seconde à la date de publication de l'œuvre originale. Cette seconde date n'est cependant pas toujours présente.
— L'URL de la page de l'**auteur** sur le site de l'Île ;
— Un champ contenant beaucoup d'informations sur l'**édition** du livre ;
— Le **lieu de publication** ;

— L'**ISBN**, dont le format semble assez variable.

— L'**éditeur** ;
— Une **collection** (par exemple, "Collection 2 continents. Série Best-sellers) ;
— Une **sous-collection**, le plus souvent uniquement identifiée par un numéro ;
— L'**année de publication**, systématiquement identique au contenu du champs "date de publication" ;
— Le **nombre de pages**
— La ou les **dimension(s)** du livre ;
— Autres informations complémentaires

Les différentes informations *peuvent* être absentes, mis à part l'éditeur et l'année de publication, systématiquement présents.

L'**ISBN** en lui-même ne nécessite pas de traitement particulier, mis à part la standardisation du format, puisque certains sont en ISBN-10, d'autres en ISBN-13. Les séparateurs habituels ne sont pas présents. Dans certains cas, il y a également la présence de séparateurs (espaces ou tirets).

ADP Les données en provenance de Messageries ADP contiennent plus de 120 000 livres, dont environ 15 000 publiés par un éditeur québécois.
Ces données ont été récupérées en ligne et sont au format ONIX for Books [1], un standard pour la représentation d'information du monde du livre sous format d'arbre XML. Chaque entité est donc constituée d'un sous-arbre contenant des propriétés imbriquées selon une arborescence contenant des sous-champs.
58 sous-champs ont été identifiés comme contenant de l'information utile. Les champs les plus importants et les plus fréquemment utilisés sont :

— *PublishingDetail*, contenant de nombreux détails d'édition, tels que :
  — L'**éditeur** (nom, site web)
  — La **date de publication**
  — Le **statut de la publication** du livre.
— *DescriptiveDetail*, contenant des informations descriptives sur le livre en soi, dont :
  — Les **contributeurs** (essentiellement le ou les **auteur(s)**)
  — Le **type** de livre (roman, bande dessinée, etc.), aini qu'une indication du schéma de classification utilisé (par exemple Thema [2])
  — L'**audience** visée en termes d'âge

1. https ://bisg.org/page/onixforbooks
2. https ://www.editeur.org/151/Thema/

— La **langue** du livre (36 en anglais, le reste en français)
— Le **titre** du livre
— Les **dimensions** du livre

Dépot Légal Bibliothèques et Archives nationales du Québec (BAnQ) a fourni l'information du Dépôt Légal sous format MARC21. Un premier fichier contient des métadonnées par rapport à des documents concernés, environ 60 000 se rapportant à la littérature québécoise. Un second contient les informations sur les personnes (autorités). Aucune clé ne lie les deux documents, mis à part le nom des personnes (auteurs principalement). Un travail de prétraitement est nécessaire pour lier les mentions d'auteurs du fichier de documents à l'entité correspondante du fichier d'autorités.
Les données du Dépôt Légal sont riches et variées, mais ont des standards de nomenclature variables et de nombreux champs non nettoyés, rendant difficile l'uniformisation des données.

Hurtubise Le jeu de données en provenance d'Hurtubise contient des données sur 1314 livres, incluant notamment :

— Trois **ISBN** par livre, pour les versions papier, PDF et *epub*
— Une division du titre de livre en **titre principal**, **sous-titre** et titre de la **série**
— Un champs contenant les **contributeurs** et leur rôles
— un **résumé**
— Une **catégorisation** du livre selon la taxonomie Thema, divisée en sujet principaux et sujets détaillés
— Des quantificateurs détaillés sur le contenu du livre (géographique, historique, langue, âge, etc.)
— Pour les quatre entrées concernant des livres scolaires, des indicateurs des **niveaux scolaires** Français et Québécois.

D'autres informations présentes incluent le **nombre de pages**, la **date de parution** ainsi que l'**éditeur** (systématiquement "Éditions Hurtubise").

L'annexe C contient un tableau détaillé des champs de tous les jeux de données sources et leur correspondance dans le modèle ontologique décrit à la section 4.3.

4.3 L'ontologie MCCQ

De l'analyse des besoins de utilisateurs découlent trois principales caractéristiques que doit présenter l'ontologie développée. L'ontologie doit :

1. Avoir la capacité de répondre aux besoins utilisateurs (questions de compétence) ;
2. Représenter toutes les données fournies ;
3. Être compatible avec les données littéraires, et idéalement culturelles, d'autres bases de connaissances.

Choix du modèle de base

Le dernier critère permet à l'ontologie développée d'être exploitée facilement par d'autres experts du domaine et la compatibilité avec des jeux de données ouverts d'autres institutions. Ainsi est motivée l'exploration des modèles existants - spécifiques au monde du livre ou concernant les données culturelles plus globalement - qui est présentée à la section 2.2. Cette considération enjoint d'écarter des modèles peu répandus tels que The Bibliographic Ontology [29].

Certains modèles d'intérêt plus répandus, tels que BibFrame [30] ou MODS/RDF avec MADS/RDF [31], suivent les standards de catalogage américains et les besoins internes de la librairie du congrès (Library of Congress). Ces modèles n'ont pas non plus les termes requis pour distinguer entre des concepts tels que l'*Expression* et la *Manifestation* (voir la prochaine section).

Les trois modèles candidats retenus sont ainsi FRBR, CIDOC-CRM (modèle à la base du projet de la cinémathèque présenté au chapitre 3) et IFLA-LRM. Ces modèles sont présentés à la section 2.2. Bien que ces trois modèles présentent les caractéristiques recherchées, notamment la compatibilité avec d'autres jeux de données ouverts et la richesse expressive, les avantages qu'IFLA-LRM possède sur les autres modèles le rend le choix évident comme base pour l'ontologie. Notamment, IFLA-LRM est le modèle successeur de FRBR, rendant ce dernier obsolète pour nos besoins, d'autant plus qu'il existe une certaine rétrocompatibilité dans LRM.

Pour ce qui est de CIDOC-CRM, ce modèle est plus étendu, complexe et généraliste que LRM, car son domaine d'application est plus vaste, mais il omet certains concepts importants en données bibliographiques, dont notamment la distinction *Expression/Manifestation* (définie à la prochaine section). À noter qu'il existe un effort conjoint entre les comités responsables du développement de IFLA-LRM et CIDOC-CRM pour intégrer ensemble ces deux modèles, rendant possible une éventuelle compatibilité.

Conséquemment, IFLA-LRM est retenu comme base de l'ontologie à développer (section 4.3.1), et des extensions s'y rajoutent pour représenter les connaissances hors de son domaine, notamment les prix littéraires (section 4.3.2).

4.3.1 Le modèle LRM

Cette sous-section présente le modèle principal développé, qui est basé sur LRM. Un modèle simplifié est ensuite présenté à la sous-section 4.3.2.

Livre L'ontologie LRM, tout comme son prédécesseur FRBR, sépare le concept de "livre" en quatre niveaux d'abstraction différents :

- *Work* : L'oeuvre abstraite, c'est à dire l'idée et les concepts du contenu sémantique de l'oeuvre.
- *Expression* : La codification de l'idée en texte ou symboles, son expression en langue.
- *Manifestation* : Description de l'objet physique publié contenant l'expression.
- *Item* : Un objet physique incarnant la manifestation, c'est-à-dire une copie tangible.

Le niveau de l'item est absent de notre modélisation, puisque le système développé ne gère pas de collections.

Un exemple de la structure des informations d'un livre est présenté à la figure 4.1. On y présente les entités requises pour la représentation d'une oeuvre conceptuelle extraite du jeu de données d'Hurtubise, l'expression qui la réalise et la manifestation dans laquelle elle est encapsulée, ainsi que les entités représentant l'éditeur et l'auteur du livre. Ces trois niveaux sont dotées de propriétés selon ce niveau d'abstraction, par exemple le titre de l'oeuvre ou encore le nombre de pages d'une manifestation.

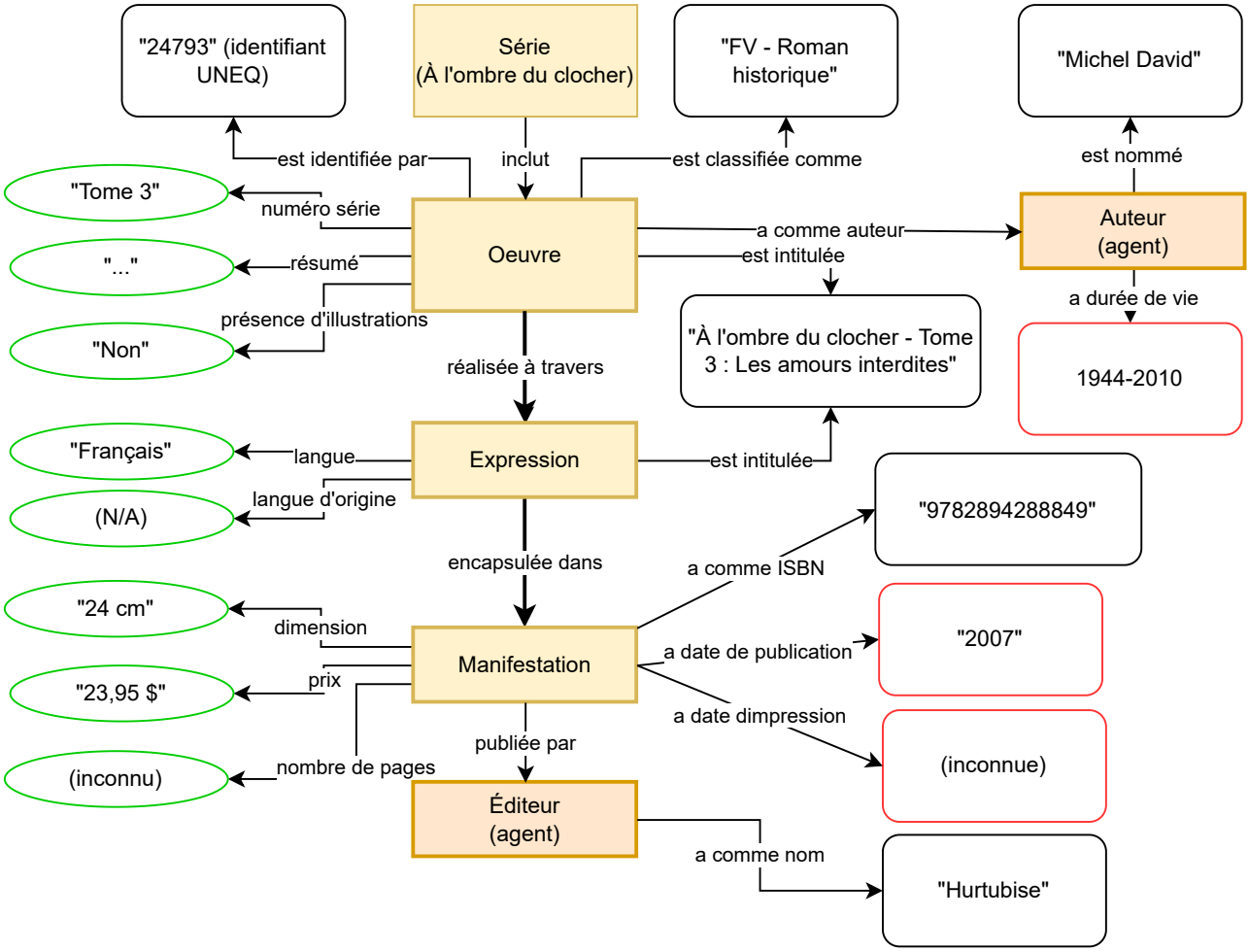

Figure 4.1 Exemple de la représentation d'un livre et entité connexes avec un modèle basé sur LRM

Nomens Les *nomens* sont une entité particulière du modèle LRM. Un *nomen* est une combinaison de sons ou symboles utilisée pour référer à une entité du monde. Chaque entité du monde peut posséder plusieurs *nomens*, mais chaque *nomen* ne réfère qu'à une entité. Des exemples de nomens incluent :

— Le nom d'un auteur : "Michel David"
— Le titre d'une oeuvre : "À l'ombre du clocher"
— Un identifiant unique, tel qu'un ISBN : "9783161484100"

Puisque les *nomens* sont des entités, il est possible de spécifier des métadonnées supplémentaires concernant l'utilisation ou l'origine d'un *nomen*, telles que sa langue, la source du standard de catalogage utilisée ou l'autorité l'ayant assigné. Ceci peut permettre notamment de varier l'affichage par langue, ou encore d'utiliser un format particulier pour le tri d'un

index. Il est également possible de spécifier une hiérarchie de *nomens*. Par exemple, "H.S. Thompson" peut être un *nomen* dérivé de "Hunter S. Thompson".

4.3.2 Extensions

Bien que la grande majorité des classes d'entités requises pour le monde du livre soient incluses dans le modèle LRM, certaines informations complémentaires pouvant aider à répondre aux besoin d'utilisateurs sont manquantes, notamment au niveau des revues et commentaires d'utilisateurs par rapport aux oeuvres. D'autres classes et propriétés ont donc du être développées pour représenter ces concepts.

De plus, le niveau de complexité de LRM, particulièrement au niveau des nomens et de la représentation des dates à travers des entités de laps de temps (time-span), rend les requêtes SPARQL plus complexes, puisqu'il existe plus d'entités intermédiaires entre l'entité du monde concernée et la valeur assignée à l'attribut souhaité. Pour faciliter l'accès aux informations le plus communes, dont les noms d'auteurs, dates de publication et titre de livres, des équivalences entre des chaînes de propriété LRM et des propriétés de données plus simples issues du modèle de schema.org ont été déclarées dans l'ontologie, profitant ainsi du régime d'inférence pour permettre l'utilisation de requêtes simplifiées.

Prix littéraires

Un des requis est la possibilité d'interroger la base de connaissances au sujet des prix littéraires remportés par des auteurs. Plus particulièrement, il est nécessaire d'être capable de représenter le fait qu'un prix a été attribué à un auteur, pour une ou plusieurs oeuvres, en une année. Puisque LRM ne contient pas les entités et propriétés requises, nous avons crée une extension basée sur une classe *Award Attribution*, dont un exemple d'utilisation est présenté à la figure 4.2.

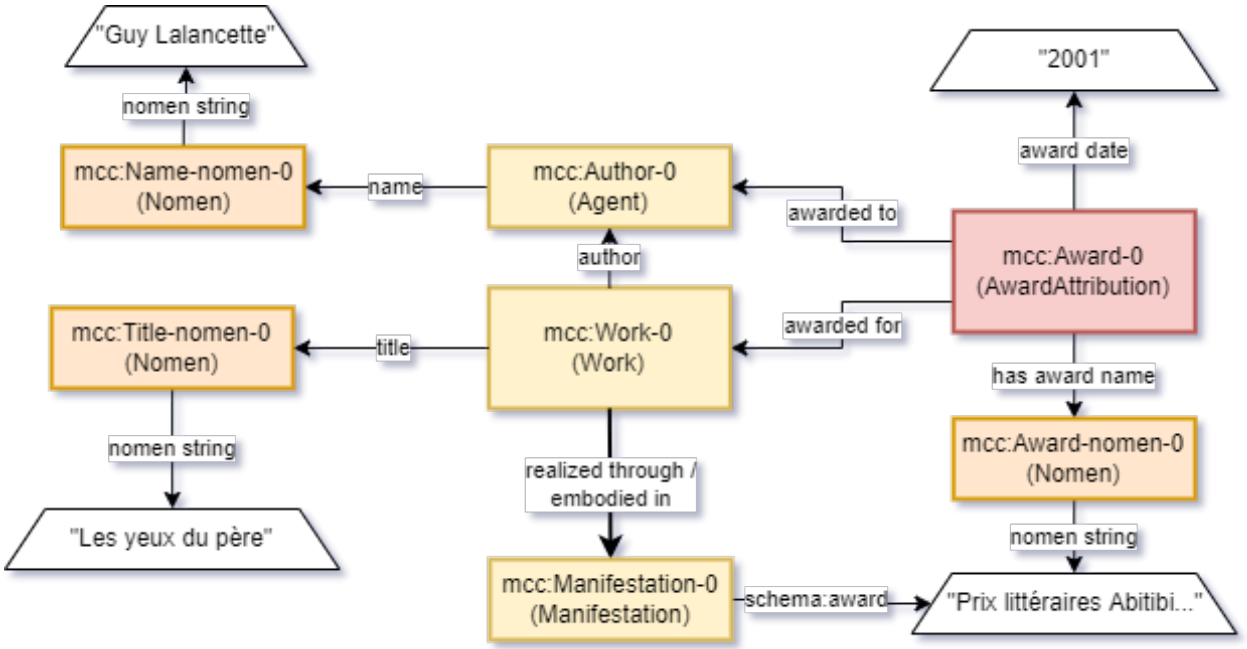

Figure 4.2 Exemple de la représentation de l'attribution d'un prix littéraire à un auteur pour une oeuvre

Ajout de sous-propriétés à LRM

Certaines relations nécessitent une représentation plus fine que celle offerte par LRM. Dans ces cas, des sous-propriétés dérivées sont ajoutées au modèle de base.

Le premier cas concerne les relations de création entre personnes et oeuvres ou expressions. Bien que la propriété LRM *was created by* permette de déclarer une relation de création, elle n'en spécifie pas la nature. Les propriétés plus explicites suivantes sont donc dérivées : *was written by*, *was illustrated by*, *was translated by*, etc.

Dans le deuxième cas, bien qu'il existe une relation *has appellation* entre les nomens et autres entités (oeuvres, auteurs, éditeurs, etc.), les nomens peuvent référer à des types de nomenclature distincts. Le type d'appellation du nomen est donc explicité par des sous-propriétés : *has complete title*, *has subtitle*, *has last name*, *has ISBN*, etc.

Mis à part les sous-relations entre entités, des sous-propriétés de données ont également été ajoutées (relations entité-attribut) pour spécifier des valeurs de différents types. Par exemple, la propriété décrivant les caractéristiques physiques d'une manifestation, *extent*, est dérivée en *dimension*, *number of pages*, *is illustrated* et *format*.

Modèle simplifié avec schema.org

Le modèle LRM possède l'avantage d'offrir de la richesse sémantique au niveau des abstractions de niveaux d'oeuvre, des appellations (nomens) et l'expression des dates. Par contre, cette richesse se traduit également par une plus grande complexité au niveau des requêtes SPARQL interrogeant le graphe. La figure 4.3 illustre la structure d'une oeuvre, reliée à son titre par une suite de deux relations : la relation au nomen, et la relation du nomen à sa valeur. Il en va de même pour les dates ; retrouver la date de publication de sa manifestation requiert le passage par l'entité intermédiaire *Time-span*.

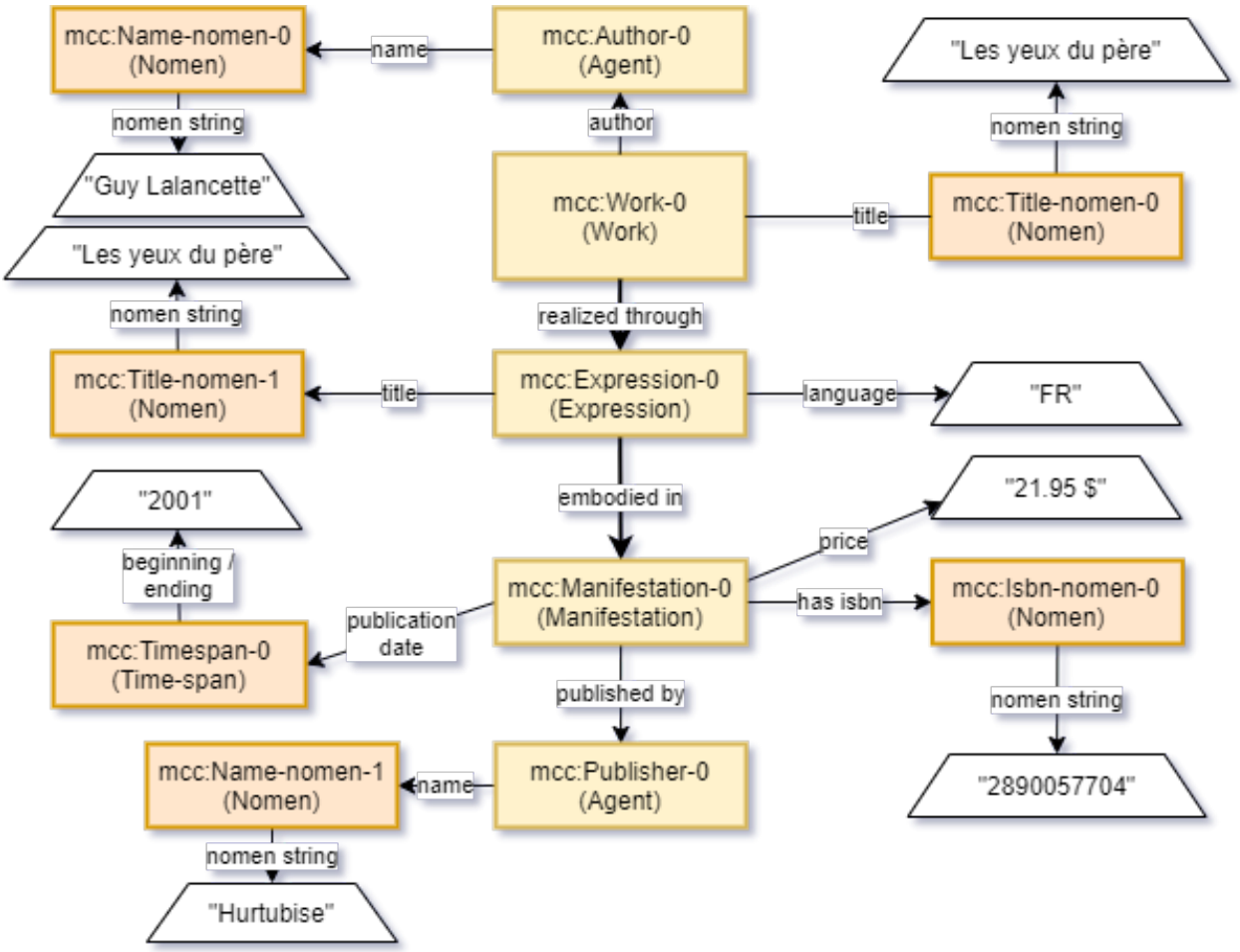

Figure 4.3 Représentation d'un livre selon LRM

Un des objectifs étant de développer une ontologie simple d'utilisation, on propose comme solution d'ajouter des correspondances entre le modèle principal basé sur LRM et un modèle plus simple, basé sur le modèle de données de schema.org, présenté précédemment à la section

2.2.

Pour ce faire, il suffit de déclarer des équivalences entre des chaînes de propriétés LRM et l'équivalent selon le vocabulaire schema.org. Ainsi, le régime d'inférence du module de requêtes SPARQL est en mesure d'inférer les relations nécessaires à partir de la représentation LRM et de répondre directement à des requêtes formatées selon le modèle simplifié basé sur schema.org. Ces raccourcis sont présentés à la figure 4.4.

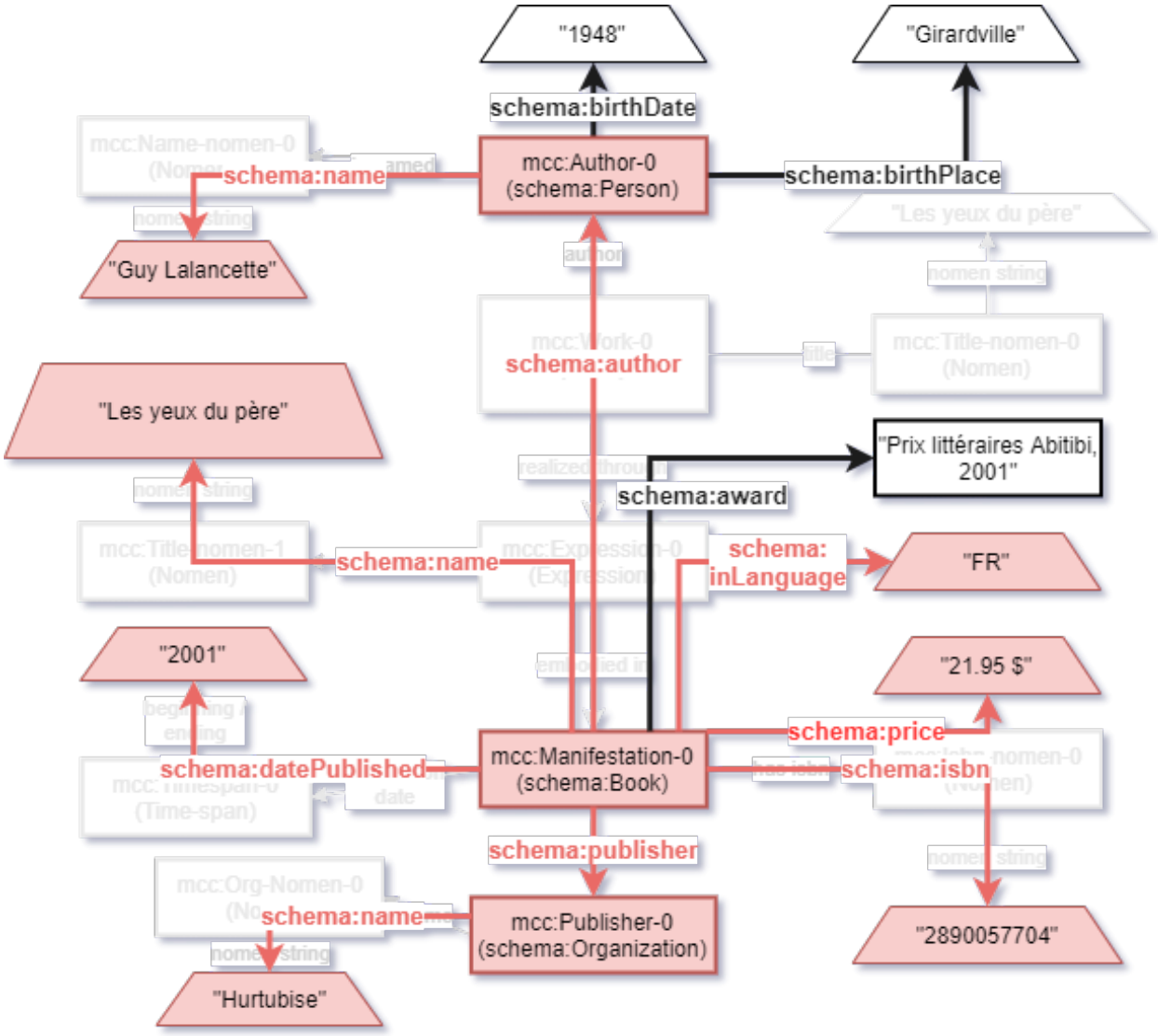

Figure 4.4 Représentation d'un livre selon Schema.org

Bien qu'utiliser le vocabulaire Schema.org rende plus facile l'accès à certains attributs, il y a également une perte de richesse. Notamment, l'abstraction de livres en plusieurs oeuvres

est absente, obligeant à répondre à la question : est-ce que diverses éditions d'une même oeuvre constituent plusieurs livres ? Si c'est le cas, alors la réponse à la requête "quels sont les livres écrits par un auteur donné" renvoie la liste de toutes les éditions et formats possibles de tous les livres de cet auteur. Si ce n'est pas le cas, alors une même oeuvre se retrouverait avec plusieurs dates de publication, plusieurs ISBN, prix, etc., sans possibilité de déterminer, par exemple, quelle date de publication devait être associée à quel ISBN.

L'utilisation conjointe des deux modèles permet donc de simplifier les requêtes de base, tout en offrant la possibilité de répondre à des demandes plus particulières.

Une fois l'ontologie modélisée, l'étape suivante, décrite à la prochaine section, est la traduction des données des formats source vers le graphe de connaissances.

4.4 Traduction des données

Le défi de cette partie du processus est la traduction de quatre jeux de données hétérogènes en un seul graphe de connaissances homogène. Plusieurs étapes consécutives sont nécessaires pour y arriver, tel qu'illustré à la figure 4.5. Ces étapes seront décrites en détail dans cette section.

4.4.1 Récupération et extraction

La première étape consiste à récupérer les données des partenaires et les stocker dans un format de travail intermédiaire. La méthode de récupération des données est variable, selon le fournisseur :

— ADP : Requêtes HTML à leur serveur, récupérant un fichier XML au format Onix.
— BAnQ : Parcours de deux fichiers XML au format MARC 21 (oeuvres et autorités)
— L'Ile : Récupération directement à partir du site litterature.org
— Hurtubise : Lecture d'un tableur Excel

Les données sont ensuite stockées dans un fichier JSON. Toutes les données sont représentées avec un ensemble de champs uniformes, mais aucun nettoyage de valeurs n'est fait à ce stade. On référera par la suite à ce format intermédiaire en JSON par l'expression **Représentation intermédiaire**. Les figures 4.6, 4.7, 4.8 et 4.9 illustrent la structure source des données, et leur représentation intermédiaire en JSON.

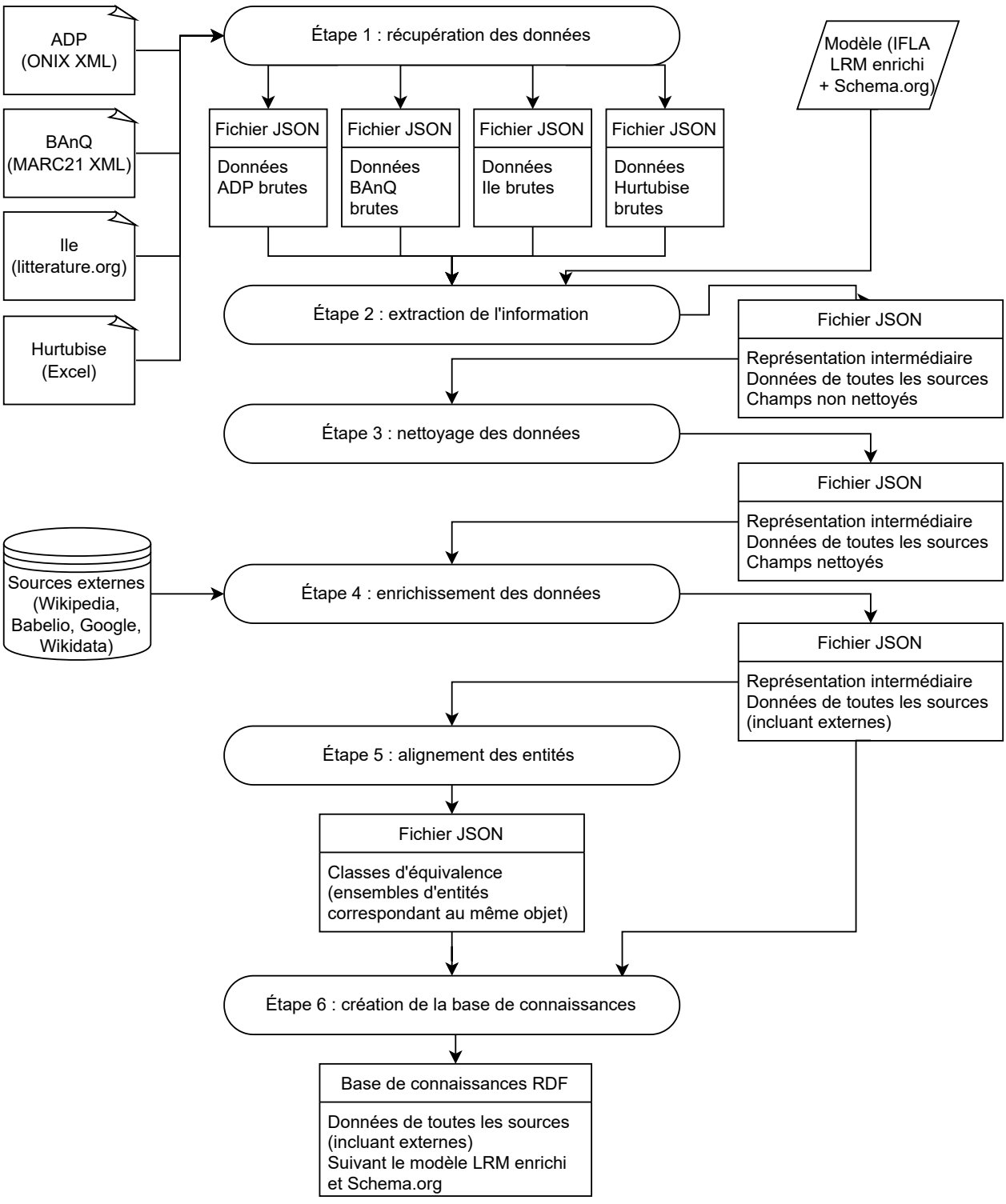

Figure 4.5 Étapes du traitement de données

À la fin de l'étape 2 du processus, l'on dispose donc d'un unique fichier de représentation

Données ADP

```
<TitleDetail>
  <TitleType>01</TitleType>
  <TitleElement>
    <TitleElementLevel>01</TitleElementLevel>
    <TitleText>Les dames de Beauchêne</TitleText>
    <Subtitle>Tome III</Subtitle>
  </TitleElement>
</TitleDetail>
```

Représentation intermédiaire

```
"titre": "Les dames de Beauchêne",
"numeroTome": "III",
"titreSerie": "Les dames de Beauchêne",
"titreComplet": "Les dames de Beauchêne, Tome III"
```

Figure 4.6 Extraction des informations de titre pour des données d'ADP

Données BAnQ

```
<ns0:datafield tag="245" ind1="1" ind2="0">
  <ns0:subfield code="a">À l'ombre du clocher :</ns0:subfield>
  <ns0:subfield code="b">[roman historique] /</ns0:subfield>
  <ns0:subfield code="c">Michel David</ns0:subfield>
</ns0:datafield>
<ns0:datafield tag="505" ind1="1" ind2=" ">
  <ns0:subfield code="a">t. 1 : Les années folles ; t. 2 : Le fils de Gabrielle. 2007</ns0:subfield>
</ns0:datafield>
```

Représentation intermédiaire

```
"titre": "À l'ombre du clocher",
"numeroTome": "1",
"titreSerie": "À l'ombre du clocher",
"sousTitre": "Les années folles"
"titreComplet": "À l'ombre du clocher, tome 1 : Les années folles"
```

Figure 4.7 Extraction des informations de titre pour des données de BAnQ

Données Hurtubise

Titre	Sous-titre	Titre de la série
À l'ombre du clocher	Tome 1	Les années folles	À l'ombre du clocher

Représentation intermédiaire

```
"titre": "À l'ombre du clocher",
"numeroTome": "1",
"titreSerie": "À l'ombre du clocher",
"sousTitre": "Les années folles"
"titreComplet": "À l'ombre du clocher, tome 1 : Les années folles"
```

Figure 4.8 Extraction des informations de titre pour des données d'Hurtubise

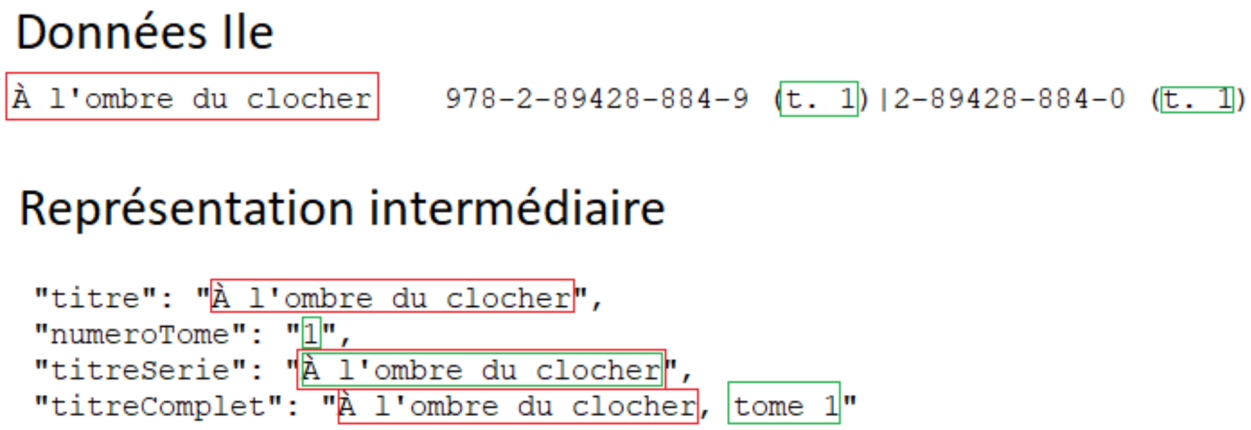

Figure 4.9 Extraction des informations de titre pour des données d'Ile

intermédiaire contenant les données de toutes les sources, sans nettoyage.

4.4.2 Nettoyage

À ce stade, la structure des données est uniforme, mais le contenu des champs ne l'est pas nécessairement. À cause des différents standards de catalogage et de propreté des données entre les sources et même à l'intérieur des sources, une même information peut être représentée différemment d'une entité à l'autre. Certaines sources indiquent le numéro de tome d'une oeuvre en chiffres romains, d'autres en chiffre arabes ; les noms des auteurs prennent parfois la forme "Nom de famille, Prénom", et parfois "Prénom Nom de famille", etc.

Ces différences nuisent à la cohérence des données, ainsi qu'à leur alignement (voir la sous-section 4.4.3), puisque, par exemple, il est difficile de déterminer si deux ISBN sont identiques si un d'entre eux contient des tirets de séparation, et l'autre non, ou encore si une année de naissance correspond à une date de naissance.

Nous avons donc analysé, pour chaque champ, les formats possibles du contenu en entrée, et déterminé le format de sortie uniforme souhaité. Certains champs, tels que les dimensions du livre, n'ont pas été nettoyés et uniformisés, possédant trop de formats différents et requérant un travail trop important pour l'envergure du projet.

À la fin de cette étape, le format du fichier de représentation intermédiaire reste identique, mais le contenu des champs a été nettoyé et uniformisé.

Désambiguïsation Tout comme dans le projet de la Cinémathèque, les lieux reviennent fréquemment comme type d'information, soit dans le contexte de lieux de naissance ou de

lieux de publication. Dans le cas du présent projet, ces lieux sont fréquemment encodés avec beaucoup de caractères erronés dans les données sources. Par exemple, on peut repérer la chaîne de caractères "[(de) Montréal]", qui fait référence à la ville de Montréal, Québec. Pour désambiguïser les lieux présents dans les données, on utilise d'abord des expressions régulières pour retirer les ensembles de caractères qui ne devraient pas y figurer.

Par la suite, il est possible d'interroger l'API Google Places [3] avec la chaîne de caractères nettoyée, qui renvoie l'identifiant unique du lieu, s'il a été identifié, ainsi que les identifiants des lieux parents. Ainsi, le résultat de la création d'entités pour les lieux ainsi récupérés et les lieux dont ils font partie est une taxonomie des lieux. Cette taxonomie enrichit le graphe et permet de répondre plus adéquatement aux questions par rapport aux lieux. Elle est structurée sous forme d'arbre, dont les noeuds proches de la racine correspondent à des pays, les noeuds intermédiaires à des régions administratives, et les feuilles à des villes, voir même des arrondissements.

Par exemple, si on avait précédemment l'information que l'auteur A était né à Montréal et l'auteur B, à Québec, il aurait été impossible de retrouver ces deux auteurs avec une requête demandant quels auteurs sont nés au Québec, puisque leurs lieux de naissances auraient été exprimés en simples chaînes de caractères. Avec une taxonomie des lieux, le graphe exprime que Montréal, tout comme Québec, fait partie de la région de Québec ; il est donc possible de répondre à la requête de manière plus précise.

4.4.3 Problème de l'alignement

L'utilité d'une base de connaissances unique regroupant des informations provenant de sources diverses repose sur l'alignement des entités provenant de ces sources. Une majeure partie de l'intérêt provient de la mise en commun de métadonnées provenant d'interprétations différentes du monde. Par contre, les différences majeures dans la qualité, la structure et les conventions des données sources rendent coûteuse ou difficile l'alignement d'entités basé sur des règles heuristiques. Les titres et noms présentent des variations de collection en collection ; les numéros de tome sont encodés sous des formats variables (chiffres romains ou arabes, dans le même champs que l'ISBN ou que le titre, etc.), plusieurs informations de types différents sont présentes dans un même champs d'édition selon la source ou même selon l'époque à laquelle une oeuvre à été cataloguée (comme illustré dans la colonne Ile du tableau C.4), et ainsi de suite. Ce problème est complexe, et son envergure justifie de lui consacrer un chapitre entier. Le chapitre 5 explique ainsi le problème de l'alignement dans le cadre d'unification de collections de données du monde de la culture québécoise. À la fin

3. https ://developers.google.com/maps/documentation/places/web-service/overview

de l'étape d'alignement, le format du fichier de représentation intermédiaire reste identique, mais on y retrouve également une liste de *classes d'équivalence* : chaque classe contient une liste d'entités qui ne représentent qu'une seule et unique entité du monde réel.

4.4.4 Génération du graphe de connaissances

Le résultat de l'étape précédente, celle de l'alignement, est une liste d'ensemble d'entités représentant une unique entité du monde réelle. Chacun de ces ensembles sera traduit en une seule entité du graphe de connaissances final. On réfère à cette entité amalgamée comme étant une *entité canonique*.

Bien que plusieurs enregistrements de sources différentes devraient théoriquement avoir des valeurs identiques pour des faits tels que la date de naissance d'une personne ou le lieu d'une publication, dans la réalité des choses, ces attributs sont fréquemment légèrement différents.

Par exemple, une source peut spécifier qu'un livre à été publié à Montréal, alors qu'une autre peut spécifier plus précisément qu'il a été publié dans l'arrondissement Saint-Laurent. Il arrive également fréquemment que les valeurs textuelles, telles qu'un titre ou un nom, comportent des imperfections qu'on ne peut identifier et corriger automatiquement.

Lorsqu'il existe plusieurs valeurs possibles pour un attribut donné d'une entité canonique, il est nécessaire d'effectuer un choix parmi ces possibilités. La méthode retenue est un classement des sources par ordre de confiance. La valeur d'un attribut de l'entité canonique est donc fixée selon celle de l'entité de la source de la plus hautement cotée possédant cet attribut. Un exemple est présenté au tableau 4.1.

Attribut	Source 1	Source 2	Source 3	Entité canonique
Confiance	3	2	1	
Nom	**Paul Dupuis**	Paule Dupuis	P. Dupuis	Paul Dupuis
Date de naissance	(absent)	**25-11-1984**	1984	25-11-1984
Lieu de naissance	**Québec, CA**	Montréal, QC, CA	(absent)	Québec, CA

Tableau 4.1 Choix des attributs d'une entité canonique

Le choix d'attributs est effectué pour toutes les entités canoniques après l'alignement. Le résultat est donc un ensemble d'entités qui devraient représenter chacune une entité du monde réel, et possédant une valeur pour chacun des attributs qu'ont les entités sources les composant.

Ces entités canoniques peuvent ensuite être traduites en RDF à l'aide d'un script Python

implémentant la conversion entre la représentation intermédiaire (RI) et le modèle final. Chacun des attributs de la RI est ainsi utilisé pour générer un ou plusieurs triplets RDF, qui sont directement insérés dans le graphe final. Grâce à la structure en graphe, cette étape est plutôt triviale. Les contraintes d'ajout d'information au graphe sont beaucoup plus simples que la création d'entités dans une base de données relationnelle. L'ajout d'une entité et de tous ses attributs au graphe n'est pas une opération atomique : chaque triplet est ajouté successivement, sans modifier les entités déjà présentes dans le graphe. Une liste exhaustive des champs utilisés de chaque source, du nom de l'attribut en représentation intermédiaire et la structure des triplets résultants pour les auteurs, oeuvres, expressions et manifestations et présentée à l'annexe C.

À la fin de cette étape, le graphe de connaissances final, exprimé en RDF, est complet, et prêt à être chargé dans un module de requêtes SPARQL pour exploitation et évaluation.

4.5 Évaluation du graphe de connaissances

Le processus d'évaluation du graphe de connaissances final est sensiblement le même que celui utilisé au chapitre 3, mis à part l'impossibilité d'effectuer de comparer le décompte des entités source au décompte des entités finales, puisque les enregistrements source peuvent être traduites en plusieurs entités de la base de connaissances et que l'étape d'alignement fusionne des entités. Un processus d'évaluation séparé, employé pour vérifier la qualité de l'alignement des entités entre sources, est présenté au chapitre 5.

La vérification de la complétude et la fonctionnalité du graphes de connaissances s'effectue à l'aide d'une évaluation qualitative de la fonctionnalité à travers les questions de compétence, et l'exploitation du graphe de connaissances par des applications prototypes l'utilisant comme source de données. On présente, dans cette section, les résultats de ces méthodes.

Réponses aux questions de compétence Mis à part les questions qui suivent, la base de connaissances est en mesure de répondre à 21 sur 29 des questions de compétence énumérées à la sous-section 4.1.1. Les questions non traitées sont les suivantes :

— 2. Où vit cet auteur ?
— 3. Ce livre-a-il été adapté ?
— 13. Quels sont les auteurs les plus lus par région ?
— 14. Dans quels pays ce livre est-il vendu ?
— 15. Ce livre a-t-il été auto-publié ou publié commercialement ?
— 21. Cet auteur a-t-il publié dans un journal ?

— 25. Combien d'exemplaires de ce livre ont été vendus en ligne à l'étranger ?
— 28. Cet auteur a-t-il participé à des entrevues / émissions ?

Mis à part la question 15, les autres questions n'ont pas de réponse dans la base de connaissances parce que les données pour y répondre ne sont ni fournies dans les ensembles sources, ni disponibles en ligne. Une bonne partie de ces informations ne serait détenue que par des entreprises privées de distribution et vente de livres, ou n'a fort probablement jamais été récoltée. Dans le cas de la question 15, il aurait fallu développer une méthode d'identification, parmi les champs d'édition, d'auto-publication, ce qui aurait demandé un travail hors de la portée du projet.

La base de connaissances est donc en mesure de répondre aux cas d'utilisation demandés mis à part ceux pour lesquels l'information est entièrement indisponible.

Prototypes exploitant la base de connaissances Un prototype de portail numérique web et un module de questions-réponses en langage naturel ont été développé par d'autres membres de l'équipe. Ces technologies utilisent comme serveur de données un module de requêtes SPARQL qui dessert la base de connaissances. Les deux prototypes ont été présentés aux acteurs du domaine ayant participé au projet, leur permettant de visualiser le contenu de la base de connaissances de manière plus conviviale. La réception des utilisateurs était très bonne, ce qui signifie que la base de connaissances est en mesure de faciliter le développement de tels outils d'accessibilité, accomplissant ainsi l'objectif de rendre ces données plus accessibles à de potentiels utilisateurs.

4.6 Conclusion

Dans ce travail, nous avons réalisé un modèle ontologique pour le monde du livre québécois, et créé une base de connaissances suivant sa structure, à partir de sources de données hétérogènes.

Le modèle réalisé consiste en réalité en deux modèles juxtaposés : un modèle simple mais limité, basé sur le vocabulaire de schema.org, et un modèle complexe, qui implémente le LRM de IFLA. Grâce à ces deux modèles, il est possible de répondre à des cas d'utilisation variés des acteurs du milieu. Le modèle simple permet d'aller chercher facilement et rapidement des informations de base sur les entités importantes du domaine. Le modèle complexe, également capable de répondre à ces questions, mais potentiellement moins convivialement, possède la richesse nécessaire pour représenter les concepts plus évolués, tel que la division des oeuvres littéraires en différentes entités de différents niveaux d'abstraction.

Par la suite, nous avons nettoyé et aligné les données de sources diverses, avant de les traduire en format de graphe suivant le modèle ontologique développé.

La capacité du modèle à répondre aux besoins des acteurs impliqués a été vérifié à l'aide de questions de compétence, ainsi que par l'utilisation de la base de connaissances comme source de données pour des prototypes d'application ; une application web et un système de question-réponses en langage naturel.

Une problématique importante, à savoir l'alignement d'entités de sources différentes, a été relevée pendant la construction des graphes RDF suivant ces modèles, et fera l'objet du chapitre 5.

CHAPITRE 5 ALIGNEMENT D'ENTITÉS DE SOURCES HÉTÉROGÈNES AVEC UN MODÈLE DE LANGUE NEURONAL

5.1 Problématique

Un problème récurrent dans la traduction de collections de jeux de données du monde culturel en bases de connaissances unifiées est l'absence d'identifiants globaux uniques pour les entités. Une méthode idéale pour aligner des entités entre sources serait l'utilisation d'identifiants tels que l'*International Standard Book Number* (ISBN), qui réfère à des éditions de livre. En pratique, par contre, la majorité des types d'entité sont seulement décrits avec des identifiants uniques à l'interne, ce qui requiert une méthode alternative pour l'alignement à travers des sources de données hétérogènes. Dans le cas de données littéraires, il est rare de trouver un type d'identifiant global unique standardisé pour des entités tels que les auteurs, même si l'ISNI (*International Standard Name Identifier*) et le VIAF (Virtual International Authority File) deviennent de plus en plus répandus. De plus, les sources de données littéraires courantes font rarement la différence entre les oeuvres conceptuelles et leurs manifestations, et, conséquemment, les oeuvres se retrouvent sans identifiants globaux uniques. L'évolution des standards de catalogage et des pratiques d'entrée de données a comme conséquence que non seulement des sources différentes utilisent des conventions différentes pour des attributs tels que les titres et les dates, mais que les sources peuvent contenir des incohérences à l'interne.

En l'absence d'identifiants, la prochaine meilleure méthode pour résoudre la duplication d'entités est la détection de représentations similaires d'entités, c'est-à-dire retrouver des entités avec des attributs suffisamment similaires pour les aligner de manière fiable. Le développement d'alignement à base de règles heuristiques a ses désavantages, notamment la nécessité de modèles homogènes et de propreté de données. L'évaluation de la similarité à l'intérieur d'une ou entre des sources est difficile, puisque certains attributs importants tels que les titres ou les noms sont formatés selon différentes règles, d'enregistrement à enregistrement, même à l'intérieur d'une même source.

L'extraction et le nettoyage des données des enregistrements avant la traduction en un modèle homogène peut aider à pallier ce problème d'"entités bruitées", permettant la comparaison d'entités prétraitées et nettoyées.

Ce processus requiert une importante quantité de travail ainsi qu'une connaissance étendue du domaine pour établir les correspondances entre champs des modèles initiaux, à cause

de facteurs tels que l'utilisation de différentes évolutions de standards de catalogage et la présence fréquente d'attributs dupliqués, d'erreurs de frappe et de valeurs mal étiquetées. Ainsi, la recherche de solutions reposant moins sur les étapes de prétraitement pourrait accélérer le développement d'un modèle d'alignement fonctionnel.

Ce chapitre explore comment les modèles de langue peuvent permettre d'unifier, en une seule base de connaissances homogène, quatre ensembles de métadonnées fournis par des acteurs variés du monde littéraire québécois. Comme les sources dont nous disposons sont principalement en français, des modèles de langue préentrainés sur des ensembles français sont utilisés pour générer des représentations de paires d'entités pour la prédiction de paires à aligner. En particulier, ce travail se penche sur l'alignement d'oeuvres et d'auteurs. On compare l'utilisation de modèles de langues à des heuristiques utilisant comme mesure la similarité de chaînes de caractères sur un ensemble pleinement nettoyé et étiqueté. Par la suite, nous évaluons les modèles de langue sur des entrées de différents formats, afin de déterminer combien de prétraitement et d'annotation sont nécessaires pour une performance de pointe. Les questions de recherche principales auxquelles on s'intéresse sont :

1. Dans le cadre de l'alignement d'entités du monde culturel, comment les modèles de langue préentrainés se comparent-ils aux règles heuristiques ?
2. Quel est l'impact du prétraitement et des formats d'entrée sur la performance de l'alignement à l'aide de modèles de langue ?

Ce chapitre est structuré comme suit. Nous présentons une vue d'ensemble des données sources à la section 5.2,et de la tâche d'alignement dans le contexte de l'architecture complète à la section 5.3.1. L'architecture des modèles de langue employés est décrite à la section 5.3.4, et les ensembles d'entraînement et la référence utilisée dans les sections 5.3.2 et 5.3.3 respectivement. Nous présentons et analysons les résultats à la section 5.4.

5.2 Présentation des données

Les jeux de données fournis décrivent des entités littéraires structurées en enregistrements. Les enregistrements consistent en des descriptions de métadonnées de livres, avec un enregistrement décrivant généralement une édition spécifique d'un livre, son contenu, sa forme physique, ses informations de publication et ses auteurs.

Les sources dont nous disposons souffrent des mêmes problèmes que les autres ensembles de données du monde littéraire quand vient le temps de les aligner. Bien que les ensembles aient des portées similaires, des différences mineures existent entre le contenu de leurs enregistrements. Un jeu peut décrire une série de livres en un seul enregistrement, comme c'est le

cas dans le jeu de données de la BAnQ, ou encore être distribué à travers plusieurs enregistrements, comme dans le cas des données d'ADP. Les conventions pour les titres et les noms varient également : certaines sources retirent les pronoms au début, certaines placent le prénom après le nom de famille pour les auteurs, d'autres scindent les titres et les sous-titres en champs différents, et ainsi de suite. Les enregistrements sont de qualité variable : certains ne contiennent que des caractères en majuscule dans leurs champs textuels, d'autres sont tronqués après un certain nombre de caractères. Les attributs que possèdent les entités varient également. Par exemple, BAnQ est la seule source qui a l'attribut de date de naissance pour les auteurs. Il est également fréquemment le cas que des attributs existant dans un modèle de données n'aient pas de valeur pour une grande portion des enregistrements. Finalement, certaines sources ont des collections séparées pour les auteurs et les livres, alors que d'autres concatènent les informations quant aux deux en un seul enregistrement.

Pour entraîner les modèles de langue à la tâche d'alignement, des jeux de données de paires d'alignements positifs (paires d'entités qui devraient être alignées) et négatifs (paires qui ne devraient pas être alignées) doivent être générés. Un identifiant global unique, l'ISBN, est disponible à travers les jeux de données pour certaines entités et peut aider à la génération de ces ensembles. La génération et le contenu de ces ensembles sont présentés à la sous-section 5.3.2 suivante.

5.3 Méthodologie

Le modèle de données final visé est celui présenté au à la section 4.3.1. L'importance de l'alignement dans le pipeline menant à la population de cette base de connaissances finale est décrit à la sous-section 5.3.1. La sous-section 5.3.2 qui explique la structure et la génération des ensembles de données étiquetés servant à l'évaluation et l'entraînement du précédent modèle. La méthode à laquelle nous comparons l'alignement à l'aide de modèles de langue, soit l'alignement heuristique, et les expérimentations et métriques, se retrouvent à la sous-section 5.3.3, alors que l'architecture du modèle d'alignement à base de modèles de langue est décrite à la sous-section 5.3.4.

5.3.1 La phase d'alignement du pipeline complet

Après l'extraction d'entités des jeux de données originaux, les entités doivent être alignées, avant leur ajout à la base de connaissances finale. Conceptuellement, la tâche est d'identifier des groupes d'entités extraites représentant la même entité du monde réel, à l'aide d'un module d'alignement qui détermine si les représentations canoniques de deux groupes d'entités

impliquent que ces groupes devraient être fusionnés. Une fois que deux groupes d'entités sont fusionnés en un nouveau groupe, ce groupe est à nouveau comparé aux autres groupes d'entités existants pour vérifier s'il ne devrait pas être fusionné à nouveau. Ce processus permet l'intégration graduelle d'entités provenant, éventuellement, de nouveaux jeux de données. Une fois que le bassin de groupes d'entités atteint un équilibre, c'est-à-dire qu'il n'est plus possible de trouver de liens entre groupes d'entités, la traduction finale des entités en une base de connaissances unifiée peut être effectuée. L'algorithme 1 décrit ce processus.

Algorithme 1 Alignement d'entités

```
E les classes d'équivalence existantes ;
O les entités non classées ;
L(x) la liste des paires d'entités composant la classe d'équivalence x ;
G(x) la représentation canonique d'une classe d'entités ;
A(x, y) alignement d'une paire de représentations canoniques ;
function ALIGNEMENT(E, O)
    p ← ∅
    for e in E do
        p ← p ∪ L(e)
    end for
    while O ≠ ∅ do
        for e in E do
            p = G(o)
            f = G(e)
            if A(p, f) then
                e ← e ∪ o
                for l in L(E) do
                    p ← p ∪ {(o, l)}
                end for
                O ← O ∪ {e}
            else
                E ← E ∪ {o}
            end if
            O ← O \ {o}
        end for
    end while
    return E
end function
```

5.3.2 Jeux de données labelisés

Des jeux de données labelisés sont requis pour les entités qui seront à aligner lors des expériences, soit pour les oeuvres et les auteurs. Les ensembles, résumés au tableau 5.1, sont constitués de paires de représentations d'entités avec comme étiquette 1 pour les paires positives et 0 pour les paires négatives.

Tableau 5.1 Description des ensembles d'entraînement, de test et de validation employés

Entité	Entraînement	Validation	Test	Positifs	Négatifs
Oeuvre	7 111	1 017	2 033	4 399	5 762
Auteur	8 661	1 082	4 331	4 700	9 374

Les paires positives pour les oeuvres sont identifiées en utilisant le seul identifiant unique qui est disponible à travers les sources : l'ISBN des manifestations associées aux oeuvres. La cardinalité des relations entre oeuvres et manifestations permet d'inférer que les oeuvres qui ont des manifestations avec des ISBNs identiques réfèrent à une seule et même oeuvre du monde réel. 3 950 paires positives ont ainsi été généré automatique. De plus, 449 paires positives ont été annotées semi-automatiquement, en retrouvant d'abord des paires d'oeuvres ayant des titres très similaires (avec un ratio de Levenshtein, défini par l'équation 5.1, entre 80 et 99) et au moins un auteur similaire (ratio de Levenshtein plus grand que 50 entre leurs noms). Ces paires candidates ont ensuite été étiquetés manuellement en paires positives et négatives.

$$Levenshtein(str1,\ str2) = \left(1 - \frac{distance\ d'édition}{max(len(str1),\ len(str2))}\right) * 100 \tag{5.1}$$

Pour ce qui est des auteurs, sans un identifiant unique (seule une source contient des identifiants ISNI ou VIAF), on doit assumer que deux enregistrements d'auteurs qui ont écrit des oeuvres identiques et ont des noms très similaires (en utilisant le ratio de Levenshtein défini par l'équation) sont identiques. En l'occurrence, les noms de deux auteurs sont considérés comme étant similaires si le ratio est égal ou supérieur à 60. Les règles utilisées pour la génération des paires positives sont les suivantes, avec les ratios de Levenshtein correspondants indiqués entre parenthèses :

1. Si deux oeuvres ont des manifestations avec des ISBN identiques, alors elles devraient être alignées.

2. Si deux auteurs ont des noms similaires (> 60) et ont écrit des oeuvres ayant des manifestations avec le même ISBN, alors ils devraient être alignées.

Pour les paires négatives, des paires d'enregistrements sont choisies au hasard. Si deux enregistrements choisis n'ont pas d'ISBN partagé et qu'il y a une grande distance d'édition entre leurs noms ou titres, soit un ratio de Levenshtein inférieur à 50 (ou autre caractéristique fondamentalement incompatible, comme des dates de naissance différentes pour des auteurs), alors elles sont considérées comme ne pouvant pas référer à la même entité du monde réel. Les règles utilisées pour la génération des paires négatives sont les suivantes, avec les ratios de Levenshtein correspondants indiqués entre parenthèses :

1. Si deux oeuvres ont des titres très différents (< 50), alors elles ne devraient pas être alignées.
2. Si deux auteurs ont des noms très différents (< 50), alors ils ne devraient pas être alignés.

Pour ajouter des exemples plus compliqués, des paires négatives d'oeuvres plus similaires ont été annotées manuellement. Parmi les oeuvres ayant des titres similaires (ratio de Levenshtein entre 80 et 99), 496 paires négatives ont été repérées.

Choix de la composition des exemples Pour éviter un déséquilibre trop important entre la quantité de paires positives et négatives, le nombre de paires négatives d'auteurs est limité au double du nombre de positives. Pour les oeuvres, le nombre de paires négatives est limité au nombre de paires positives. Les paires positives et négatives sont générées, puis séparées en ensembles d'entraînement, validation et test contenant des proportions identiques de paires positives et négatives.

Comme les ISBNs ont été utilisés comme clés de jointure pour les ensembles d'entraînement, de validation et de test, ces identifiants ne peuvent être inclus dans l'encodage en chaîne de caractères des entités. Comme le but est d'entraîner un modèle qui aligne en l'absence d'identifiants uniques partagés, inclure l'ISBN pourrait fausser les résultats, puisque les paires positives ne seraient probablement alignées qu'avec cette information.

5.3.3 Métriques d'évaluation et résultats de référence

Notre *baseline* de méthode d'alignement est basé sur la similarité de chaînes de caractères entre attributs appartenant aux entités. N'utiliser que les noms des entités pourrait être problématique, puisqu'il existe des paires d'entités qui ne sont pas à aligner qui ont des noms

plus similaires que certaines paires d'entité qui sont à aligner. Ceci est illustré à la figure 5.1, qui trace un estimé de densité des ratios de Levenshtein entre les titres d'oeuvres à aligner ou non. Le graphique de gauche montre une vue globale de ces ratios, alors que celui à droite illustre le recouvrement entre les similarités des paires positives et négatives. Ce recouvrement indique qu'il n'est pas possible de se fier uniquement au ratio de Levenshtein pour aligner des oeuvres selon leurs titres entre des valeurs de 60 et 80 sur 100.

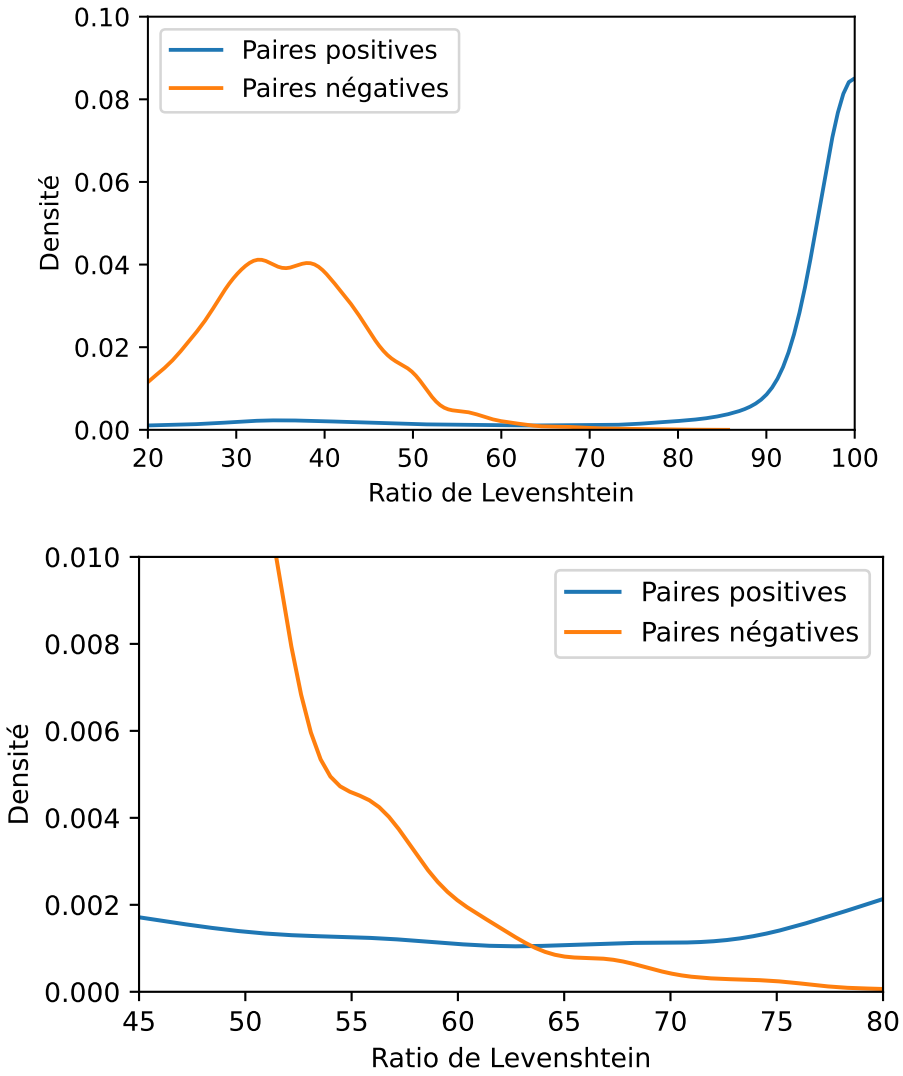

Figure 5.1 Estimés de densité des ratios de Levenshtein de titres d'oeuvres identiques (paires positive) et distinctes (paires négatives) utilisant le format d'entrée 1 (Tableau 5.4)

Se fier uniquement aux titres créerait des faux positifs ; comparer les noms des auteurs de chaque paire candidate élimine ces cas. Si deux oeuvres ont des noms d'auteurs très similaires en plus d'avoir des titres similaires, alors la méthode heuristique considère qu'elles

sont à aligner. De même, les auteurs sont alignés s'ils ont des noms similaires et ont écrit des oeuvres avec des noms similaires.

La méthode heuristique employée utilise donc un ensemble de règles statiques pour déterminer si deux entités sont à aligner ou non. Un premier ensemble de règles sert à détecter les paires d'entités candidates à l'alignement, alors qu'un deuxième ensemble élimine les paires candidates contenant des incohérences.

Règles heuristiques Les listes suivantes correspondent aux règles d'alignement pour les oeuvres et les auteurs. Les attributs doivent avoir une similarité correspondante au ratio de Levenshtein donné entre parenthèses. Une valeur de 100 indique que deux attributs doivent être identiques.

Les règles de similarité d'attributs pour aligner les oeuvres sont les suivantes :

1. Titres similaires (50) et numéro de tome identiques (100)
2. Titres similaires (50) et noms d'auteurs similaires (70)
3. Titres similaires (50) et noms d'éditeurs similaires (70)
4. Titres similaires (60) et nombres de pages identiques (100)
5. Titres similaires (80) et dates de publication identiques (100)
6. Titres similaires (92) et lieu de publications similaires (90)
7. Titres similaires (95)

Dès qu'une seule de ces règles est satisfaite, une paire d'oeuvres est alignée par la méthode heuristique.

Les règles heuristiques employées pour trouver les paires d'entités *auteur* candidates sont les suivantes :

1. Noms similaires (40) et titres d'oeuvres écrites similaires (50)
2. Noms similaires (40) et années de naissance identiques (100)
3. Noms similaires (40) et années de décès identiques (100)
4. Noms similaires (70) et langues d'oeuvres écrites identiques (100)
5. Noms similaires (75)

Dès qu'une seule de ces règles est satisfaite, une paire d'auteur devient candidate à l'alignement. Par la suite, une paire d'auteurs candidate doit satisfaire les deux règles suivantes, faute de quoi elle n'est pas alignée :

1. Si les deux auteurs ont des années de naissance, elles doivent être identiques.
2. Si les deux auteurs ont des années de décès, elles doivent être identiques.

Les métriques employées pour l'évaluation des règles heuristiques et des modèles de langue neuronaux sont le rappel (équation 5.2), la précision (équation 5.3), l'exactitude (équation 5.4) et le score F1 (équation 5.5). Les valeurs possibles pour ces métriques sont entre 0 et 1, et les variables utilisées dans ces métriques sont expliquées au tableau 5.2.

$$Rappel = \frac{VP}{VP + FN} \tag{5.2}$$

$$Précision = \frac{VP}{VP + FP} \tag{5.3}$$

$$Exactitude = \frac{VP + VN}{VP + VN + FP + FN} \tag{5.4}$$

$$F1 = 2 * \frac{Précision * Rappel}{Précision + Rappel} \tag{5.5}$$

Variable	Explication
VP	Vrai positif : la méthode a aligné une paire qui devait être alignée
VN	Vrai négatif : la méthode n'a pas aligné une paire qui ne devait pas être alignée
FP	Faux positif : la méthode a aligné une paire qui ne devait pas être alignée
FN	Faux négatif : la méthode n'a pas aligné une paire qui devait être alignée

Tableau 5.2 Variables utilisées dans le calcul des métriques d'évaluation

Les résultats des heuristiques et des modèles de langue neuronaux pour ces métriques sont présentés à la section 5.4 et analysés à la section 5.5.

5.3.4 Architecture des modèles de langue

L'architecture utilisée est inspirée par les travaux récents [25] et [24], mais diffère par la nature des données (données provenant de bases de données relationnelles, données structurées), la langue (français) et la nature des expériences menées, soit sur l'impact des formats d'entrée et stratégies de prétraitement. Les modèles de langue utilisés sont couplés avec une couche de neurones pleinement connectée et une couche de classification SoftMax, ajoutées

à la sortie. L'entraînement et l'évaluation sont effectués sur les ensembles d'entraînement, de validation et de test présentés à la sous-section 5.3.2. Les modèles préentrainés choisis pour la génération de représentations vectorielles à partir de séquences représentant des paires d'entités sont CamemBERT [23] et FlauBERT [1], qui sont les deux basés sur l'architecture de RoBERTa [32]. Ces modèles génèrent des représentations vectorielles en nombre réels à partir de texte. Ces vecteurs sont ensuite classifiées par la couche SoftMax comme étant à aligner (étiquette 1) ou non (étiquette 0). Les hyperparamètres trouvés en [24] sont réutilisés pour la taille des groupe (*batchsize*), soit 32, et le taux d'apprentissage (3e-5).

Format des données d'entrée Les modèles de langue à la BERT [3], tels que ceux utilisés dans le présent travail, prennent comme entrée du texte, composé d'une ou plusieurs séquences, et, pour les tâches de classification, une étiquette à prédire. Dans le présent cas, chaque entrée représente une paire d'entités, composée de deux chaînes de caractères représentant chacune une entité, et une étiquette qui identifie si ces entités devraient être ou non alignées. Les jeux de données étiquetés contiennent une entrée par ligne, avec un caractère de tabulation séparant chacun des trois éléments : les chaînes de caractères pour les entités 1 et 2, et l'étiquette.

Des expériences sont menées avec quatre formats d'entrée, présentés au tableau 5.3, qui diffèrent par le prétraitement et les stratégies d'annotation employés pour générer les représentations textuelles des entités. Ces formats d'entrée concatènent les valeurs des attributs extraits en chaînes de caractères pour ces représentations.

Tableau 5.3 Formats d'entrée pour différentes stratégies de prétraitement et annotation

	Pré-traitement et annotation	Structure
1	Avec nettoyage RegEx et noms d'attributs LRM, valeurs et annotations colonne/valeur	[C] attribut1 [V] valeur1 [C] attribut2 [V] valeur2 [...] [C] attributN [V] valeurN
2	Sans nettoyage, avec noms originaux, valeurs et annotations colonne/valeur	[C] attribut1 [V] valeur1 [C] attribut2 [V] valeur2 [...] [C] attributN [V] valeurN
3	Sans nettoyage ni annotations colonne/valeur, avec noms originaux et valeurs	attribut1 valeur1 attribut2 valeur2 [...] attributN valeurN
4	Valeurs sans nettoyage, annotations colonne/valeur ni noms	valeur1 valeur2 [...] valeurN

Les stratégies d'annotation peuvent ajouter des étiquettes correspondant à des noms

d'attributs, et des étiquettes spéciales, par exemple "[C]", qui aident à baliser quelles portions des segments réfèrent à des étiquettes ou des valeurs.

Le premier format est basé sur ce qui était proposé par Li et al, 2020 [25]. Les données utilisées pour ce format sont nettoyées avec des expressions régulières qui normalisent la ponctuation, retirent des caractères spéciaux superflus et restructurent les chaînes de caractères pour qu'elles soient de formats identiques (par exemple, tous les noms sont écrits sous la forme "Nom, Prénom") à travers les sources de données. Les données sont restructurées pour que leurs attributs respectent le modèle LRM final visé. Les représentations des entités consistent en une alternance de noms d'attributs et de valeurs d'attributs, séparés par des étiquettes spéciales qui indiquent si la portion suivante de la chaîne de caractères est un nom d'attribut ([C]) ou sa valeur ([V]. Les noms d'attributs sont d'une longueur de un ou deux caractères : t pour titre, st pour sous-titre, a pour nom d'auteur, etc.

Le second format est similaire, conservant l'étiquetage, mais ne faisant ni restructuration de données, ni nettoyage de valeurs. Les noms des attributs peuvent conséquemment être des chaînes textuelles ou des identifiants de champs standardisés, tel qu'un code MARC21 [1], selon la source.

Le troisième format ne fait ni restructuration de données, ni nettoyage de valeurs, ni étiquetage de portion de chaîne. Les noms des attributs dépendent de la source.

Le format final n'utilise aucun nettoyage, aucun étiquetage de sous-chaîne et ne contient pas de noms d'attributs ; les représentations des entités consistent simplement en une concaténation des valeurs de leurs attributs.

Des exemples pour chacun de ces formats d'entrée pour deux mêmes entités à aligner sont montrés au tableau 5.4.

1. https ://www.loc.gov/marc/bibliographic/

Tableau 5.4 Exemples pour les formats d'entrée pour un alignement qui devrait être positif. La première séquence provient de BAnQ, et la deuxième d'Hurtubise.

#		Valeur
1	Seq 1 :	**[C] t [V]** Être un héros **[C] e [V]** La Courte échelle **[C] st [V]** des histoires de gars **[C] lp [V]** Montréal **[C] np [V]** 218
	Seq 2 :	**[C] t [V]** Être un héros **[C] ap [V]** 2011 **[C] a [V]** Simon Boulerice **[C] e [V]** La Courte échelle **[C] st [V]** des histoires de gars **[C] lp [V]** Montréal **[C] np [V]** 218
2	Seq 1 :	**[C] 245a [V]** Être un héros : **[C] 245b [V]** des histoires de gars / **[C] 260b [V]** La Courte échelle, **[C] 300a [V]** 1 ressource en ligne (218 p.) : **[C] 260a [V]** Montréal :
	Seq 2 :	**[C] 0 [V]** Être un héros : des histoires de gars **[C] 2 [V]** Boulerice, Simon **[C] 3 [V]** La Courte échelle, 2011, 218 p. **[C] 1 [V]** 2011 **[C] 4 [V]** Montréal
3	Seq 1 :	**245a** Être un héros : **245b** des histoires de gars / **260b** La Courte échelle, **300a** 1 ressource en ligne (218 p.) : **260a** Montréal :
	Seq 2 :	**0** Être un héros : des histoires de gars **2** Boulerice, Simon **3** La Courte échelle, 2011, 218 p. **1** 2011 **4** Montréal
4	Seq 1 :	Être un héros : des histoires de gars / La Courte échelle, 1 ressource en ligne (218 p.) : Montréal :
	Seq 2 :	Être un héros : des histoires de gars Boulerice, Simon La Courte échelle, 2011, 218 p. 2011 Montréal

La section 5.4 compare les résultats obtenus par les modèles de langue sur l'ensemble de test à ceux obtenus avec la méthode heuristique.

5.4 Résultats

Dans cette section, les résultats expérimentaux pour les questions de recherche sont présentés et discutés. Dans les tableaux de résultats, le plus haut score pour chaque métrique pour chaque entité est indiqué en gras.

Le première expérience effectuée s'intéressait à la capacité des modèles de langue à effectuer la tâche d'alignement. Pour ce faire, une implémentation de l'alignement à l'aide de ML a été entraînée et évaluée sur des données nettoyées, reformatées selon le modèle LRM et annotées avec des étiquettes spéciales (format 1 du tableau 5.4). Pour évaluer le potentiel de cette méthode, nous comparons, au tableau 5.5, ses résultats contre ceux obtenus sur le même ensemble d'évaluation avec les heuristiques.

Type de modèle	Entité	F1	Rappel	Exactitude	Précision
Heuristiques	Auteur	0.9972	0.9986	0.9982	0.9959
	Oeuvre	0.8739	0.9375	0.8829	0.8185
EN-BERT unc.	Auteur	**0.9983**	**0.9993**	**0.9988**	**0.9972**
	Oeuvre	**0.9835**	**0.9841**	**0.9857**	**0.9830**

Tableau 5.5 Comparaison entre les performances des ML et heuristiques pour l'alignement

Ces résultats montrent que les modèles de langue en combinaison avec un classificateur SoftMax surpassent les heuristiques développées avec connaissance du domaine. Ces résultats démontrent également que la tâche d'alignement d'auteurs est triviale en comparaison avec la tâche d'alignement d'oeuvres ; un ensemble de test plus difficile serait requis pour une meilleur comparaison. Par contre, les meilleures heuristiques sont incapables d'égaler la performance quasi-parfaite des modèles de langue pour les auteurs, avec les modèles de langue ayant plus de capacité à bien aligner les cas limites.

Entraîner des modèles séparés pour des types d'entité différents utilise plus de ressources. Une expérience menée vise à tester si un modèle simple entraîné sur une tâche conjointe d'alignement d'auteurs et d'oeuvres peut atteindre des résultats similaires de performance. Le tableau 5.6 démontre qu'un modèle simple est aussi efficace que des modèles séparés.

Type de modèle	F1	Rappel	Exactitude	Précision
Heuristiques	0.9493	0.9755	0.9620	0.9246
EN-BERT unc.	**0.9946**	**0.9940**	**0.9961**	**0.9953**

Tableau 5.6 Résultats du modèle conjoint Auteurs-Oeuvres

Le tableau 5.7 compare la performance de modèles préentraînés anglais, multilingues et français. Même si les données employées sont en français, les modèles préentraînés en français ne semble pas avoir d'avantage important sur les autres.

Type de modèle	F1	Rappel	Exactitude	Précision
EN-BERT unc.	**0.9946**	0.9940	**0.9961**	**0.9953**
M-BERT	**0.9946**	0.9961	**0.9961**	0.9931
M-DistilBERT	0.9929	0.9927	0.9948	0.9931
CamemBERT	0.9925	0.9953	0.9945	0.9897
FlauBERT	0.9931	**0.9966**	0.9940	0.9897

Tableau 5.7 Comparaison de résultats entre modèle de langue

Le modèle le plus performant de la dernière expérience, soit BERT anglais sans casse [2], est ensuite utilisé pour les quatre formats d'entrée présentés à la section 5.4. Les résultats sur ces différents formats sont affichés au tableau 5.8.

Format d'entrée	F1	Rappel	Exactitude	Précision
1	**0.9946**	0.9940	**0.9961**	0.9953
2	0.9940	**0.9957**	0.9956	0.9923
3	**0.9946**	0.9927	**0.9961**	**0.9965**
4	0.9940	0.9953	0.9956	0.9927

Tableau 5.8 Comparaison des résultats du meilleur modèle sur différents formats d'entrée

Étant donnés les résultats similaires ou même plus élevés en terme de score F1 lorsque peu ou pas de prétraitement est employé, on conclut que ces étapes coûteuses peuvent être omises sans risque significatif pour la performance. Plus d'étapes de prétraitement requièrent le développement de plus de règles spécifiques au domaine avec peu de réutilisabilité et qui requièrent de la maintenance en cas d'ajout de plus de sources de données. Les résultats pour le format 4 démontrent que les modèles de langue permettent d'éviter ce problème pour l'alignement d'entités, dans le contexte des jeux de donnés utilisés. On conclut que, du moins dans notre cas spécifique et contrairement à ce qui a été trouvé dans le travail de Li et al. [25], l'ajout de balises spéciales `[C]` et `[V]` n'a pas d'impact important sur la performance, comme l'illustre la différence négligeable entre les résultats des formats 2 (balisé) et 3 (non balisé), qui utilisent les mêmes noms et valeurs d'attributs.

5.5 Conclusion

Nous avons proposé un modèle d'alignement d'entités basé sur des modèles de langue, pour l'alignement d'enregistrement digitaux du monde de la culture exprimés en français. Les résultats démontrent que de tels modèles à base de transformeurs préentraînés sur des tâches de langage masqué sont des outils puissants pour l'alignement d'entités du monde culturel. On tire également comme conclusion que des performances élevées peuvent être obtenues même en n'entraînant que sur des ensembles limités, sans prétraitement et très hétérogènes de données labelisés.

Même si les méthodes heuristiques obtiennent de bonnes performances, leurs coûts relativement élevés en temps de développement et leur manque de réutilisabilité font de l'alignement à l'aide de modèles de langue une alternative compétitive en mesure d'apprendre

2. https ://huggingface.co/bert-base-uncased

automatiquement les règles d'alignement nécessaires. L'utilisation de données hétérogènes a peu d'importance sur la performance des modèles de langue dans les expériences menées, ce qui peut faciliter l'intégration de davantage d'ensembles de données.

De plus, les résultats montrent que le prétraitement n'améliore pas de manière significative la performance de l'alignement à base de modèles de langue, ce qui permettrait d'éviter plus de temps de développement. La différence de performance avec des ensembles de données prétraitées et étiquetées comme proposé par Li et al. [25] est minimales ; l'étiquetage des valeurs avec les noms d'attributs et étiquettes spéciales n'a pas d'impact majeur selon les expériences menées dans le présent travail.

CHAPITRE 6 CONCLUSION

6.1 Synthèse des travaux

Ce travail se penche sur la modélisation et la création de bases de connaissances pour le monde de la culture québécoise. L'objectif était de créer des modèles ontologiques adaptés aux mondes québécois du cinéma et de la littérature, et de peupler les bases de connaissances correspondantes avec des ensembles de métadonnées existants. Ces bases de connaissances devraient être compatibles avec d'autres bases de connaissances distribuées publiquement.

D'abord, nous avons établi au chapitre 2 un portrait des ontologies pouvant représenter les connaissances de ces domaines, qu'elles soient généralistes ou plus spécifiquement adaptées aux métadonnées culturelles. Ainsi, nous avons pu développer des modèles compatibles avec ceux utilisés par d'autres initiatives similaires.

Nous avons implémenté, au chapitre 3, un modèle pour le monde du cinéma québécois, en mesure de satisfaire les cas d'utilisation proposés par les acteurs intéressés par son exploitation. Nous avons réutilisé et étendu un modèle existant, FRBRoo, et avons construit un graphe de connaissances selon le format de ce modèle, avec des données fournies par la Cinémathèque québécoise. Des questions de compétence et un prototype d'application web ont permis de présenter aux parties intéressées le contenu et les capacités de la base de connaissances réalisée.

Au chapitre 4, nous avons présenté un travail similaire de modélisation pour le monde du livre québécois, cette fois-ci avec un modèle basé sur le LRM de IFLA. Le modèle a été validé à travers des questions de compétence issues des requis des acteurs du domaine, et la base de connaissances a servi de source de données pour un prototype d'application web ainsi qu'un système de question-réponses en langage naturel. La différence majeure avec le précédent travail fut la source des données : au lieu d'un ensemble unique et standardisé, les données sources provinrent d'une variété de sources hétérogènes. Conséquemment, un travail d'alignement d'entités entre ces sources était nécessaire. Les caractéristiques des ensembles sources ont rendu difficile l'utilisation de méthodes d'alignement à l'aide d'identifiants uniques.

Au chapitre 5, avons donc étudié le potentiel des modèles de langue à base de transformeurs pour la tâche d'alignement. Au moment de réaliser ce projet, seulement deux travaux existants avaient utilisé cette technologie pour l'alignement d'entités. Nous avons créé des ensembles d'entraînement et d'évaluation pour déterminer si ces techniques seraient également applicables à notre cas d'utilisation, à savoir l'alignement d'entités du monde de la culture,

en langue française. Nos expériences ont déterminé que ces technologies étaient en mesure d'aligner de manière beaucoup plus précise et efficace les entités de nos sources hétérogènes que des techniques de comparaison plus traditionnelles.

6.2 Limitations de la solution proposée et pistes d'amélioration

Nous relevons ici certaines des limites du travail effectué, et proposons des pistes de solutions pour y pallier, ainsi que pour améliorer et étendre les bases de connaissances créées.

Base de connaissances du monde du cinéma La principale lacune actuelle du travail effectué pour modéliser l'ontologie du monde du cinéma québécois est l'utilisation de FRBR (plus spécifiquement FRBRoo) comme modèle de base. Puisqu'IFLA LRM remplace maintenant ce modèle, il est rendu obsolète. De ce fait, le modèle ontologique développé n'est plus à jour avec l'état de l'art des modèles ontologiques du monde de la culture. Il faudrait donc restructurer le modèle et la base de connaissances pour respecter le modèle LRM. Une autre amélioration possible est l'agrandissement du modèle. Dans sa version actuelle, il ne couvre qu'une portion restreinte du jeu de données source. D'autres informations existantes dans la base de données initiale pourraient être représentés dans la base de connaissances finale. Par exemple, plusieurs informations sont disponibles sur l'aspect technique de la réalisation de films, tel que les spécifications du type d'équipement utilisé pour le tournage.

Base de connaissances du monde du livre Bien qu'un travail important de nettoyage de données ait été effectué pour tenter d'améliorer la qualité des données provenant des ensembles fournis par les partenaires du projet, il ne fut pas possible d'effectuer un travail complet à cet égard à cause de l'étendue du projet et des ressources en temps disponible. Une étude plus approfondie des ensembles fournis pourrait permettre de repérer une partie plus importante des inconsistances que l'on souhaiterait corriger ou retirer des données finales. Certaines des questions de compétence se trouvent également sans réponse à cause d'un manque de données sources nécessaires. Y répondre nécessiterait l'ajout de données de nature plus commerciale, tels que des données de vente de livres détenues par un distributeur. Il faudrait également étendre le modèle pour y ajouter les classes et propriétés nécessaires pour répondre à ces questions.

Alignement d'entités à l'aide de modèles de langue L'obstacle le plus important à la réalisation de cette tâche est la génération d'ensembles d'entraînement, de validation et de test pour les modèles de langue. Une quantité importante d'exemples étiquetés est

nécessaire, et la qualité de ces exemples (en difficulté, et en précision des étiquettes) doit être élevée pour entraîner les modèles de manière optimale et pour que les résultats obtenus soient représentatifs de la performance des modèles. Bien qu'un mélange de techniques automatiques et semi-automatiques ait permis la génération d'ensembles assez imposants, ceux-ci ne contiennent pas nécessairement des exemples étiquetés plus difficiles à évaluer automatiquement. Il serait donc nécessaire d'investir une quantité de temps importante à étiqueter des paires d'entités à aligner ou non. Une autre amélioration possible est le point de comparaison utilisé pour évaluer la performance des modèles de langue pour ces tâches. Dans notre travail, nous avons utilisé une méthode basé sur des règles heuristiques comme référence. Cependant, d'autres méthodes d'alignement neuronales existantes, tel que celle proposée par Ebraheem et al. [33], qui utilise des réseaux neuronaux récurrents (RNN), pourraient également servir pour l'alignement d'entités du monde culturel québécois.

RÉFÉRENCES

[1] H. Le *et al.*, "Flaubert : Unsupervised language model pre-training for french," *arXiv preprint arXiv :1912.05372*, 2019.

[2] V. Sanh *et al.*, "DistilBERT, a distilled version of BERT : smaller, faster, cheaper and lighter," *arXiv preprint arXiv :1910.01108*, 2019.

[3] Google Research, "google-research/bert/multilingual.md," Oct 2019. [En ligne]. Disponible : https://github.com/google-research/bert/blob/master/multilingual.md

[4] E. Miller, "An introduction to the resource description framework." *D-lib Magazine*, 1998.

[5] C. Bizer, T. Heath et T. Berners-Lee, "Linked data : The story so far," dans *Semantic services, interoperability and web applications : emerging concepts*. IGI global, 2011, p. 205–227.

[6] G. O. Consortium, "The Gene Ontology (GO) database and informatics resource," *Nucleic acids research*, vol. 32, n°. suppl_1, p. D258–D261, 2004.

[7] S. Auer *et al.*, "Dbpedia : A nucleus for a web of open data," dans *The semantic web*. Springer, 2007, p. 722–735.

[8] R. V. Guha, D. Brickley et S. Macbeth, "Schema. org : evolution of structured data on the web," *Communications of the ACM*, vol. 59, n°. 2, p. 44–51, 2016.

[9] D. C. M. Initiative *et al.*, "Dublin core metadata element set, version 1.1," 2012.

[10] D. Soergel, "The art and architecture thesaurus (AAT) : A critical appraisal," *Visual Resources*, vol. 10, n°. 4, p. 369–400, 1995.

[11] IFLA Study Group on the Functional Requirements for Bibliographic Records , "Functional requirements for bibliographic records - final report," févr. 2009. [En ligne]. Disponible : https://www.ifla.org/files/assets/cataloguing/frbr/frbr_2008.pdf

[12] M. L. inc., "(CIDOC CRM 7.1.1) Definition of the CIDOC Conceptual Reference Model," avr. 2021. [En ligne]. Disponible : https://www.cidoc-crm.org/sites/default/files/cidoc_crm_v.7.1.1_0.pdf

[13] M. Doerr *et al.*, "FRBRoo, a conceptual model for performing arts," dans *2008 Annual Conference of CIDOC, Athens*, 2008, p. 15–18.

[14] K. Wildenhaus, "The Possibilities of Constructing Linked Data for Art Exhibition Histories," *Art Documentation : Journal of the Art Libraries Society of North America*, vol. 38, n°. 1, p. 22–34, 2019.

[15] V. A. Carriero *et al.*, "ArCo : The Italian cultural heritage knowledge graph," dans *International Semantic Web Conference*. Springer, 2019, p. 36–52.

[16] "A Creative Works Ontology for the Film and Television Industry," 2018. [En ligne]. Disponible : https://movielabs.com/distribution-specs/creative-works-ontology/

[17] P. Riva, P. Le Bœuf et M. Žumer, "IFLA Library Reference Model," *A Conceptual Model for Bibliographic Information. Hg. v. IFLA International Federation of Library Associations and institutions. Online verfügbar unter https ://www. ifla. org/files/assets/cataloguing/frbr-lrm/ifla-lrm-august-2017_rev201712. pdf*, 2017.

[18] S. Peroni, "SAMOD : an agile methodology for the development of ontologies," dans *Proceedings of the 13th OWL : Experiences and Directions Workshop and 5th OWL reasoner evaluation workshop (OWLED-ORE 2016)*, 2016, p. 1–14.

[19] A. Gillioz *et al.*, "Overview of the Transformer-based Models for NLP Tasks," dans *2020 15th Conference on Computer Science and Information Systems (FedCSIS)*, 2020, p. 179–183.

[20] J. Devlin *et al.*, "BERT : Pre-training of Deep Bidirectional Transformers for Language Understanding," 2018, cite arxiv :1810.04805Comment : 13 pages. [En ligne]. Disponible : http://arxiv.org/abs/1810.04805

[21] Y. Kim et A. M. Rush, "Sequence-level knowledge distillation," *arXiv preprint arXiv :1606.07947*, 2016.

[22] Y. Liu *et al.*, "RoBERTa : A robustly optimized BERT pretraining approach," *arXiv preprint arXiv :1907.11692*, 2019.

[23] L. Martin *et al.*, "Camembert : a tasty french language model," *arXiv preprint arXiv :1911.03894*, 2019.

[24] U. Brunner et K. Stockinger, "Entity matching with transformer architectures-a step forward in data integration," dans *International Conference on Extending Database Technology, Copenhagen, 30 March-2 April 2020*. OpenProceedings, 2020.

[25] Y. Li *et al.*, "Deep entity matching with pre-trained language models," *arXiv preprint arXiv :2004.00584*, 2020.

[26] S. Mudgal *et al.*, "Deep learning for entity matching : A design space exploration," dans *Proceedings of the 2018 International Conference on Management of Data*. VLDB Endowment, 2018, p. 19–34.

[27] Z. Yang *et al.*, "Xlnet : Generalized autoregressive pretraining for language understanding," *Advances in neural information processing systems*, vol. 32, 2019.

[28] J. Raad et C. Cruz, "A survey on ontology evaluation methods," dans *Proceedings of the International Conference on Knowledge Engineering and Ontology Development, part of the 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management*, 2015.

[29] F. Giasson *et al.*, "Bibliographic ontology," Technical report, Rapport technique, 2008.

[30] BibFrame : Bibliographic framework initiative. [En ligne]. Disponible : https://www.loc.gov/bibframe/

[31] MODS RDF initiatives.

[32] Y. Liu *et al.*, "RoBERTa : A Robustly Optimized BERT Pretraining Approach," *CoRR*, vol. abs/1907.11692, 2019. [En ligne]. Disponible : http://arxiv.org/abs/1907.11692

[33] M. Ebraheem *et al.*, "DeepER–Deep Entity Resolution," *arXiv preprint arXiv :1710.00597*, 2017.

ANNEXE A IMPLÉMENTATION DES QUESTIONS DE COMPÉTENCE POUR LE PROJET DE LA CINÉMATHÈQUE

1a La requête SPARQL de la figure A.1 permet de générer toutes les combinaisons de deux personnes ayant travaillé ensemble, le nombre de films dans lesquels elles ont collaboré, ainsi que leurs titres. Elle cherche toute instance de deux personnes occupant une fonction sur un même film, et regroupe les résultats par groupes de deux personnes. Une liste des films sur lesquels ces personnes ont collaboré et le décompte du nombre de films est retournée, en plus des noms des personnes dans un groupe. Les résultats sont ensuite triés par ordre décroissant de collaborations. Les cinq premiers résultats sont présentés à la figure A.2. La figure A.3 présente la même requête mais pour les groupes de femmes québécoises ayant le plus souvent travaillé ensemble.

```
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX cmtq: <https://data.cinematheque.qc.ca/data#>

select ?nom1 ?nom2 (count(distinct ?filmo) as ?compte)
(group_concat(distinct ?titre; separator=", ") as ?films)
where {
    ?filmo a cmtq:Filmo .

    ?participant1 a cmtq:Personne .
    ?filmo ?fct1 ?participant1 .

    ?participant2 a cmtq:Personne .
    ?filmo ?fct2 ?participant2 .

    ?filmo cmtq:a_titre_lit ?titre .
    ?participant1 rdfs:label ?nom1 .
    ?participant2 rdfs:label ?nom2 .
    FILTER(?nom1 < ?nom2) .
}  group by ?participant1 ?participant2 ?nom1 ?nom2 order by
desc(count(distinct ?filmo)) LIMIT 100
```

Figure A.1 Requête sur le modèle simple pour la question de compétence 1a

	nom1	nom2	compte
1	André Larin	Michel Bissonnette	"219"^^xsd:integer
2	André Larin	Vincent Leduc	"214"^^xsd:integer
3	Michel Bissonnette	Vincent Leduc	"214"^^xsd:integer
4	Guy Villeneuve	Michel St-Cyr	"161"^^xsd:integer
5	Edwin S. Porter	Thomas A. Edison	"156"^^xsd:integer
6	Jacquelin Bouchard	Sylvie Desrochers	"133"^^xsd:integer
7	Gilbert Rozon	Sylvie Arbour	"129"^^xsd:integer

Figure A.2 Sept premiers résultats pour la question de compétence 1a. Les titres ont été retirés pour la lisibilité

```
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX cmtq: <https://data.cinematheque.qc.ca/data#>

PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
select ?nom1 ?nom2 (count(distinct ?filmo) as ?compte)
(group_concat(distinct ?titre; separator=", ") as ?films) where
{
    ?filmo a cmtq:Filmo .

    ?participant1 a cmtq:Personne .
    ?participant1 cmtq:a_genre cmtq:Feminin .
    ?participant1 cmtq:quebecois "true"^^xsd:boolean .
    ?filmo ?fct1 ?participant1 .

    ?participant2 a cmtq:Personne .
    ?participant2 cmtq:a_genre cmtq:Feminin .
    ?participant2 cmtq:quebecois "true"^^xsd:boolean .
    ?filmo ?fct2 ?participant2 .

    ?filmo cmtq:a_titre_lit ?titre .
    ?participant1 rdfs:label ?nom1 .
    ?participant2 rdfs:label ?nom2 .
    FILTER(?nom1 < ?nom2) .
}  group by ?participant1 ?participant2 ?nom1 ?nom2 order by
desc(count(distinct ?filmo)) LIMIT 100
```

Figure A.3 Femmes québécoises ayant le plus souvent travaillé ensemble

1b La requête SPARQL de la figure A.4 permet de générer la liste de toutes les personnes avec qui Geneviève Bujold a travaillé, en ordre décroissant du nombre de collaborations. C'est une requête similaire à la requête de la figure A.1, à la majeure différence que l'identité d'une des deux personnes des « groupes » cherchés est fixé à Geneviève Bujold. On retrouve ainsi ses collaborateurs les plus fréquents. Les résultats sont présentés à la figure A.5.

```
PREFIX cmtq: <https://data.cinematheque.qc.ca/data#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>

select ?nomCollegue (count(distinct ?filmo) as ?nombreFilms)
 (group_concat(distinct ?nomFilm; separator=", ") as ?films)
 ?genreCollegue where {
    # Geneviève Bujold
    cmtq:Personne24742 rdfs:label ?nom.

    ?filmo ?fct1 cmtq:Personne24742 .
    ?filmo a cmtq:Filmo .
    ?filmo rdfs:label ?nomFilm .
    ?filmo ?fct2 ?collegue .
    ?collegue a cmtq:Personne .

    ?collegue rdfs:label ?nomCollegue.
    FILTER(?collegue != cmtq:Personne24742)
    ?collegue cmtq:a_genre ?genreCollegue
} group by ?collegue ?nomCollegue ?genreCollegue order by
desc(count(distinct ?filmo)) limit 100
```

Figure A.4 Requête sur le modèle simple pour la question de compétence 1b

	nomCollegue	nombreFilms	films	genreCollegue
1	Michel Brault	"10"^^xsd:integer	KAMOURASKA, LES NOCES DE PAPIER, MON AMIE MAX, ENTRE LA MER ET L'EAU DOUCE, LA FLEUR DE L'ÂGE, MARIE-CHRISTINE, L' EMPRISE, CLAUDE JUTRA, PORTRAIT SUR FILM, ROULI-ROULANT	cmtq:Masculin
2	Claude Jutra	"6"^^xsd:integer	KAMOURASKA, ENTRE LA MER ET L'EAU DOUCE, MARIE-CHRISTINE, CLAUDE JUTRA, PORTRAIT SUR FILM, ROULI-ROULANT	cmtq:Masculin
3	Paul Almond	"6"^^xsd:integer	THE DANCE GOES ON, ISABEL, THE ACT OF THE HEART, JOURNEY, FINAL ASSIGNMENT, THE PUPPET CARAVAN	cmtq:Masculin
4	Werner Nold	"4"^^xsd:integer	ENTRE LA MER ET L'EAU DOUCE, LA FLEUR DE L'ÂGE, MARIE-CHRISTINE, ROULI-ROULANT	cmtq:Masculin

Figure A.5 Quatre premiers résultats pour la question de compétence 1b

2a La requête de la figure A.6 retourne toutes les fonctions que Woody Allen a occupé sur des films, ainsi que les titres de ces films, classés en ordre chronologique de parution. Les résultats sont présentés à la figure A.7.

```
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX cmtq: <https://data.cinematheque.qc.ca/data#>
PREFIX frbroo:
<http://iflastandards.info/ns/fr/frbr/frbroo/>
PREFIX crm: <http://www.cidoc-crm.org/cidoc-crm/>

select ?titre (group_concat(distinct ?labelFct; separator=",
") as ?fonction) ?as where {
    # Woody Allen
    cmtq:Personne19445 a cmtq:Personne .
    cmtq:Personne19445 rdfs:label ?nom .

    ?filmo a cmtq:Filmo .
    ?filmo cmtq:a_titre_lit ?titre .
    ?filmo ?fct cmtq:Personne19445 .
    ?fct rdfs:label ?labelFct .
    ?filmo cmtq:a_annee_lit ?as .
} group by ?nom ?titre ?as order by asc(?as)  LIMIT 10000
```

Figure A.6 Requête sur le modèle simple pour la question de compétence 2a

	titre	fonction	as
1	WHAT'S NEW PUSSYCAT?	Scénario	"1965"^^xsd:short
2	WHAT'S UP TIGER LILY?	Réalisation	"1966"^^xsd:short
3	CASINO ROYALE	Interprétation	"1967"^^xsd:short
4	TAKE THE MONEY AND RUN	Réalisation, Interprétation, Scénario	"1969"^^xsd:short
5	BANANAS	Réalisation	"1971"^^xsd:short

Figure A.7 Cinq premiers résultats pour la question de compétence 2a

3 La requête de la figure A.8 cherche les personnes provenant des données de la Cinémathèque sur Wikidata, en lançant un appel à leur service de requêtes. Le service de requête permet d'envoyer une portion de la requête à un serveur distant, en l'occurrence celui de Wikidata. Par la suite, le résultat est intégré à la portion locale de la requête. Si la personne est trouvée sur Wikidata, on peut retrouver son genre. Les résultats sont présentés à la figure A.9.

```
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX cmtq: <https://data.cinematheque.qc.ca/data#>
PREFIX wdt: <http://www.wikidata.org/prop/direct/>
PREFIX wd:  <http://www.wikidata.org/entity/>
select distinct ?label ?genre where {
    ?personne a cmtq:Personne.
    ?personne rdfs:label ?label .
    BIND(STRLANG(?label, "fr") as ?labelFrancais)
    SERVICE <https://query.wikidata.org/sparql> {
        [  rdfs:label ?labelFrancais ;
           wdt:P31 wd:Q5 ;
           wdt:P21 [ rdfs:label ?genre]  ] .
        FILTER(lang(?genre) = "fr")
    }

}  limit 10
```

Figure A.8 Requête d'interrogation de Wikidata pour la question de compétence 3

	label	genre
1	Woody Allen	"masculin"@fr
2	Irwin Winkler	"masculin"@fr
3	Pedro Almodóvar	"masculin"@fr
4	Roy London	"masculin"@fr
5	Lewis Gilbert	"masculin"@fr
6	Arnaud Desplechin	"masculin"@fr
7	Chris Columbus	"masculin"@fr
8	Paul Mazursky	"masculin"@fr
9	Spike Lee	"masculin"@fr
10	Marcel Carné	"masculin"@fr

Figure A.9 Dix premiers résultats pour la question de compétence 3

La requête de la figure A.10 permet d'obtenir les dates de naissance ainsi que les identifiants VIAF des femmes de la base de connaissance de la Cinémathèque à partir des données de Wikidata. Les résultats sont présentés à la figure A.11.

```
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX cmtq: <https://data.cinematheque.qc.ca/data#>
PREFIX wdt: <http://www.wikidata.org/prop/direct/>
PREFIX wd:  <http://www.wikidata.org/entity/>
select distinct ?label ?dateNaissance ?identifiantViaf where {
    ?personne a cmtq:Personne.
    ?personne rdfs:label ?label .
    BIND(STRLANG(?label, "fr") as ?labelFrancais)
    SERVICE <https://query.wikidata.org/sparql> {
        [  rdfs:label ?labelFrancais ;
           wdt:P31 wd:Q5 ;
           wdt:P21 wd:Q6581072  ;
           wdt:P569 ?dateNaissance ;
           wdt:P214 ?identifiantViaf ] .
    }

}  limit 10
```

Figure A.10 Extraction des genres, identifiants VIAF et dates de naissances à partir de Wikidata

	label	dateNaissance	identifiantViaf
1	Pina Bausch	"1940-07-27T00:00:00Z"^^xsd:dateTime	79040630
2	Josiane Balasko	"1950-04-15T00:00:00Z"^^xsd:dateTime	162301891
3	María Novaro	"1950-09-11T00:00:00Z"^^xsd:dateTime	65097060
4	Brigitte Roüan	"1946-09-28T00:00:00Z"^^xsd:dateTime	17434214
5	Valérie Stroh	"1958-08-11T00:00:00Z"^^xsd:dateTime	46952424
6	Kathryn Bigelow	"1951-11-27T00:00:00Z"^^xsd:dateTime	27256983
7	Jocelyn Moorhouse	"1960-09-04T00:00:00Z"^^xsd:dateTime	102672348
8	Martha Coolidge	"1946-08-17T00:00:00Z"^^xsd:dateTime	24803156
9	Susanne Bier	"1960-04-15T00:00:00Z"^^xsd:dateTime	119927378
10	Agnès Varda	"1928-05-30T00:00:00Z"^^xsd:dateTime	84256688

Figure A.11 Dix premiers résultats pour la requête A.10

La requête de la figure A.12 cherche les Filmos de la base de données de la Cinémathèque sur Wikidata en utilisant leur identifiant de la cinémathèque. On sait que la propriété P4276 représente l'identifiant de Cinémathèque sur Wikidata. En trouvant les entités sujettes à cette propriété, on retrouve les identifiant Wikidata correspondant à ceux de la Cinémathèque. Les résultats sont présentés à la figure A.13

```
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX cmtq: <https://data.cinematheque.qc.ca/data#>
PREFIX wdt: <http://www.wikidata.org/prop/direct/>
PREFIX wd:  <http://www.wikidata.org/entity/>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
select *  where {
    ?film a cmtq:Filmo.
    ?film rdfs:label ?label .
    BIND(REPLACE(STR(?film),
'https://data.cinematheque.qc.ca/data#Filmo', '') AS
?idCq) .
    SERVICE <https://query.wikidata.org/sparql> {
          ?wikiId wdt:P4276 ?idCq .
    }

}  limit 10
```

Figure A.12 Extraction des genres, identifiants VIAF et dates de naissances à partir de Wikidata

	label	dateNaissance	identifiantViaf
1	Pina Bausch	"1940-07-27T00:00:00Z"^^xsd:dateTime	79040630
2	Josiane Balasko	"1950-04-15T00:00:00Z"^^xsd:dateTime	162301891
3	María Novaro	"1950-09-11T00:00:00Z"^^xsd:dateTime	65097060
4	Brigitte Roüan	"1946-09-28T00:00:00Z"^^xsd:dateTime	17434214
5	Valérie Stroh	"1958-08-11T00:00:00Z"^^xsd:dateTime	46952424
6	Kathryn Bigelow	"1951-11-27T00:00:00Z"^^xsd:dateTime	27256983
7	Jocelyn Moorhouse	"1960-09-04T00:00:00Z"^^xsd:dateTime	102672348
8	Martha Coolidge	"1946-08-17T00:00:00Z"^^xsd:dateTime	24803156
9	Susanne Bier	"1960-04-15T00:00:00Z"^^xsd:dateTime	119927378
10	Agnès Varda	"1928-05-30T00:00:00Z"^^xsd:dateTime	84256688

Figure A.13 Dix premiers résultats pour la requête A.12

ANNEXE B DEMANDES DES PARTENAIRES DU PROJET DU MONDE DU LIVRE

Les demandes suivantes ont été reproduites telles quelles à partir du document de travail fourni en début du projet. À noter que certaines demandes n'ont pu être remplies par manque d'information. Les demandes sont rassemblées par partenaire

Questions soumises par l'Institut Canadien de Québec

- Lieu/Région de résidence/de naissance des écrivain.e.s (partiellement traité : pas de lieux de résidence dans nos données)
- Anniversaire des événements marquants dans le milieu littéraire (anniversaire de naissance, de décès, de publication d'un livre marquant, etc.) (partiellement traité : pas d'information sur les livres « marquants »)
- Référents géographiques (pays, ville, quartier, commerce, etc.) ou temporels (année, événement historique important, etc.) dans les oeuvres
- Les références littéraires/musicales/culturelles présentes dans les oeuvres
- Thématiques principales des livres pour développer des événements sur un sujet précis ou des mises en valeur particulières
- Les adaptations réalisées avec les livres (théâtre, cinéma, etc.)
- Pour les livres jeunesse, précision sur le public visé, les formats des livres, la présence ou non d'illustrations, etc.
- Les prix littéraires obtenus par les oeuvres (finalistes, lauréats)
- ET SURTOUT, croiser toutes ces données : les auteurs de Québec qui ont remporté un prix littéraire, etc.

Questions soumises par BAnQ

- Les auteurs par régions, lieux de naissance et ou de résidence.
- Les auteurs féminins et masculins.
- Les auteurs les plus lus par régions.
- Des thématiques par éditeurs.
- Dans quels pays se retrouvent les livres des auteurs québécois.
- Les illustrateurs par publics cibles.
- Tous les titres publiés en gros caractères, format poche, numériques, etc.
- Les livres par nbre de pages.
- Les livres avec illustrations.

— Les livres par tranche de prix.
— Les auteurs qui ont gagné des prix.
— Les livres d'auteurs étrangers mais publiés au Québec.
— Et toutes les questions auxquelles peuvent déjà répondre les catalogues de bibliothèques.

Questions soumises par l'UNEQ

— Identifier les gens qui publient un livre pour la première fois au Québec par année
— Distinguer à coup sûr l'édition à compte d'éditeur, à compte d'auteur et l'autoédition au cours de la dernière année
— Avoir, pour un écrivain donné, la liste complète de ses publications dans des revues littéraires et d'autres périodiques
— La seule liste que nous avons trouvé qui puisse aider notre recrutement, c'est une compilation mensuelle des livres québécois à paraître par BAnQ :
https ://www.banq.qc.ca/ressources_en_ligne/livres_quebecois_paraitre.html
Mais dans ces compilations, il y a des traductions de livres étrangers, des rééditions, etc. Il faut trier la liste à la mitaine pour dégager les primoromanciers, les nouveautés, etc.
— Savoir combien de gens publient actuellement de la poésie au Québec. Ou du roman de science-fiction. Ou des livres de cuisine.
— L'organisation d'activités littéraires, à l'UNEQ, est souvent difficile à faire parce que nous ne disposons pas d'outils de recherche très précis. On ne compte plus les fois où on nous a posé des questions du genre : "Pour l'organisation d'une table ronde, ça nous prendrait un auteur d'essais de moins de 40 ans qui habite la région de Québec ; où est-ce qu'on peut trouver ça ?" Avec notre CRM, nous pouvons obtenir la liste de tous les membres de la région de Québec, on connaît aussi l'âge, mais on ne peut pas avoir une liste d'essayistes de moins de 40 ans ET qui habitent la région de Québec. Et il n'existe pas, hors de notre CRM, une liste exhaustive d'essayistes de moins de 40 ans dans la région de Québec.
— Enfin, il y a un énorme enjeu (qui dépasse l'UNEQ, c'est un enjeu pour la chaîne du livre au complet) : il est actuellement impossible de connaître l'ampleur des achats de livres électroniques hors Québec. Les statistiques de Gaspard et de l'Institut de la statistique du Québec ne comptabilisent que les ventes de livres électroniques par des détaillants québécois. Ces ventes sont généralement assez maigres (sauf depuis la mi-mars, avec la pandémie). Mais beaucoup de Québécois achètent sur Amazon et d'autres plateformes étrangères. C'est de l'argent qui est dépensé à l'étranger et qui

ne reviendra jamais ici, et on n'a aucune donnée là-dessus.

— Serait-il possible d'obtenir des données sur les livres électroniques québécois dont les ventes dépassent par exemple 1000 exemplaires dans l'année ?
— Connaitre les auteurs qui ont publié des livres numériques, et des livres audios au Québec ces 5 dernières années.
— Connaître les auteurs qui ont publié des livres dont le sujet est en lien avec des pratiques numériques ces 5 dernières années.
— Connaître tous les auteurs qui ont été invités à intervenir/témoigner lors d'émissions et/ou articles de presse sur Radio Canada ces 12 derniers mois
— Souvent nous recherchons aussi par thématique, exemple :
  — Auteurs ayant écrit des livres parlant d'amour chez les personnes agées
  — Auteurs ayant écrit de la fiction sur les changements climatiques

ANNEXE C CORRESPONDANCE ENTRE LES CHAMPS DES DONNÉES ET LES TRIPLETS RDF RÉSULTANTS POUR LES DONNÉES DU PROJET DU MONDE DU LIVRE

Tableau C.1 Correspondances entre les champs des données et les triplets RDF résultants pour les auteurs

ADP	BAnQ	Ile	Hurtubise	Représentation intermédiaire	Triplet(s) (A = Auteur)
DescriptiveDetail/ Contributor/ ContributorRole[A01]/ PersonNameInverted/	100/a, 600/a, 700/a, 100/a, 378/q	Auteur, Nom	Contributeurs, Contributeur (premier)	nomComplet	`<A mcco:MCC-R13-5-1 N>` `<N mcco:LRM-E9-A2 nomComplet>` `<A schema:name nomComplet>`
	100/d, 600/d, 700/d, 024/d			dateNaissance	`<A schema:birthDate dateNaissance>`
	100/d, 600/d, 700/d, 024/d			dateDeces	`<A schema:deathDate dateDeces>`
	100/e, 600/e, 700/e		Contributeurs	rolesOeuvres, rolesExpressions	`<O mcco:MCC-R5-1 A>` `<E mcco:MCC-R6-[1..9] A>`
	375/a			genre	`<A schema:gender genre>`
	370/a			lieuNaissance	`<A schema:birthPlace lieuNaissance>`
	370/b			lieuDeces	`<A schema:deathPlace lieuDeces>`
	370/a			nationalite	`<A schema:nationality nationalite>`
		Spécialité		genreLitteraire	`<A mcco:MCC-E6-A2-1 genreLitteraire>`
		Biographie		biographie	`<A mcco:MCC-E1-A2-1 biographie>`

Tableau C.2 Correspondances entre les champs des données et les triplets RDF résultants pour les oeuvres et expressions

ADP	BAnQ	Ile	Hurtubise	Représentation intermédiaire	Triplet(s) (O = Oeuvre, E = Expression)
DescriptiveDetail/ TitleDetail/ TitleElement/ TitleElementLevel[01]/ TitleText/	240/a, 240b, 245/a	Titre	Titre	titre	<O mcco:MCC-R13-2-3 NU> <NU mcco:LRM-E9-A2 titre> <O schema:name titre>
DescriptiveDetail/ Collection/ CollectionTYpe[10]/ TitleDetail/ TitleText/	440, 490, 830	Titre	Titre de la série	titreSerie	<S mcco:LRM-R18 O> <S mcco:MCC-R13-2-3 NU> <NU mcco:LRM-E9-A2 titreSerie> <S schema:name titreSerie>
DescriptiveDetail/ TitleDetail/ TitleElement/ TitleElementLevel[01/ Subtitle/	245/b	Titre	Sous-Titre	titreSerie	<O mcco:MCC-R13-2-4 NST> <NST mcco:LRM-E9-A2 sousTitre>
DescriptiveDetail/ Collection/ CollectionType[10]/ TitleDetail/ PartNumber/	20/a	Titre, ISBN	Titre	numeroTome	<O mcco:MCC-R13-2-5 NIT> <NIT mcco:LRM-E9-A2 numeroTome>
DescriptiveDetail/ AudienceRange/	521/a		Quantificateur d'âge	publicCible	<O mcco:LRM-E3-A3 publicCible>
DescriptiveDetail/ Language/	41/a, 41/h		Langue, Langue Origine	langue (oeuvre), langue (expression)	<O mcco:MCC-E2-A2-3 langue> <E mcco:LRM-E3-A6 langue>

Tableau C.3 Correspondances entre les champs des données et les triplets RDF résultants pour les oeuvres et expressions (suite)

ADP	BAnQ	Ile	Hurtubise	Représentation intermédiaire	Triplet(s) (O = Oeuvre, E = Expression)
DescriptiveDetail/ Subject/ MainSubject/ SubjectScheme Identifier[93]/ SubjectCode/			Sujet THEMA principal, Sujet THEMA	categorie/thema	a <O mcco:themaCategory TC> <TC mcco:qualifier thema>
			Quantificateur géoraphique (sic) Quantificateur de langue, Quantificateur Historique, Niveau soclaire (sic) FR, Niveau scolaire QC, Quantificateur d'intérêt, Quantificateur d'âge, Quantificateur de style	categorie/thema (divers)	<O mcco:themaCategory TC> <TC mcco:qualifier thema>
	082/2, 082/a			categorie/dewey	<O mcco:deweyCategory DC> <DC mcco:qualifier dewey>
	586/a			prixLitteraire	<PL mcco:MCC-R37 O> <PL mcco:MCC-R13-6 NP> <NP mcco:LRM-E9-A2 prixLitteraire[nom]> <PL schema:name prixLitteraire[nom]> <PL mcco:MCC-R35-4 DP> <DP mcco:LRM-E11-A1 prixLitteraire[date]> <DP mcco:LRM-E11-A2 prixLitteraire[date]>

Tableau C.4 Correspondances entre les champs des données et les triplets RDF résultants pour les manifestations

ADP	BAnQ	Ile	Hurtubise	Représentation intermédiaire	Triplet(s) (M = Manifestation)
ProductIdentifier/ ProductIDType[15]/ IDValue	020/a	ISBN	ISBN Papier, PDF, epub	isbn	<M mcco:MCC-R13-3-1 N> <N mcco:LRM-E9-A2 isbn> <M schema:isbn isbn>
Extent/ExtentType[07] ExtentValue/	300/a	Informations d'édition	Nombre de pages	nombrePages	<M mcco:MCC-E4-A2-2 nombrePages> <M schema:numberOfPages nombrePages>
	300/d	Informations d'édition		dimensions	<M mcco:MCC-E4-A2-1 dimensions> <M schema:size dimensions>
PublishingDetails/ PublishingDate/ PublishingDateRole[01] Date/	260/c	Date, Informations d'édition	Date de parution, Année de parution	datePub, anneePub	<M mcco:MCC-R35-2 T> <T mcco:LRM-E11-A1 datePub> <T mcco:LRM-E11-A2 datePub> <M schema:datePublished datePub>
PublishingDetails/ CityOfPublication	260/a	Lieu		lieuPub	<M mcco:MCC-R35-4 lieuPub>
PublishingDetails/ SupplyDetail/Price	020/c			prix, monnaie	<M mcco:MCC-E4-A5-1 prix + monnaie>
DescriptiveDetail/ ProductFormDetail	020/a, 020/b, 020/q	ISBN	ISBN Papier, PDF, epub	format	<M mcco:MCC-E4-A2-4 format>
PublishingDetails/ Publisher/ PublishingRole[01]/ PublisherName	260/b	Informations d'édition	Éditeur	editeur	<M mcco:MCC-R7-1 E> <M schema:publisher E> <E mcco:LRM-R13 N> <N mcco:LRM-E9-A2 editeur>
			Résumé	resume	<M mcco:LRM-E1-A2 resume> <M schema:abstract resume>