Un problème d'identification d'entités nommées dans des bases de donnés documentaires

Michel Chein, Alain Gutierrez, Michel Leclère

Un problème d'identification d'entités nommées dans des bases de donnés documentaires

M. Chein
GraphIK - LIRMM
University of Montpellier, France
michel.chein@lirmm.fr

A. Gutierrez
GraphIK - LIRMM
University of Montpellier, France
alain.gutierrez@lirmm.fr

M. Leclère
GraphIK - LIRMM
University of Montpellier, France
michel.leclere@lirmm.fr

ABSTRACT

Cet article concerne la construction, la vérification et la réparation de liens d'égalité et de différence entre entités nommées dans des bases de données documentaires. Nous proposons une méthode générale que nous précisons dans le cas où les entités nommées sont des personnes. Le problème étudié peut être considéré comme un cas simple d'un problème fondamental du web sémantique celui de la construction ou de la vérification de liens *owl :sameAs* et *owl :differentFrom* entre identifiants représentant des entités du monde extérieur. Un prototype a été construit et a été évalué sur la base sudoc qui est le catalogue collectif de l'enseignement supérieur géré par l'Agence Bibliographique de l'Enseignement Supérieur.

Keywords

Coréférence, réconciliation de références, identification d'entités nommées, bases de données documentaires, représentation de connaissances et raisonnements

1. INTRODUCTION

L'objet de cet article est le "contrôle" du *lien de référence* qui est implicitement établi entre un identifiant d'un système d'information et l'entité nommée du monde réel que cet identifiant référence au sein du système. Dans le système, on utilise des identifiants (c'est-à-dire des chaînes de caractères) pour désigner ces entités du monde réel que ce système souhaite référencer. Pour résoudre un tel problème il est nécessaire de représenter des connaissances et selon le langage de représentation des connaissances utilisé, ces identifiants seront nommés des URI (en RDF), des constantes (en logique des prédicats du premier ordre), des individus (en logique de description)...

Le problème fondamental que pose un tel système d'information est celui de l'identification de l'entité nommée représentée par un identifiant, et son dual celui du choix de l'identifiant pour désigner une entité nommée. Pour se débarrasser de ce problème, les formalismes de représentation des connaissances font souvent l'hypothèse de nom unique (UNA pour "Unique name Assumption"). L'hypothèse UNA signifie que deux identifiants différents désignent deux entités différentes du monde représenté par le système. Cette hypothèse est généralement supposée satisfaite dans le domaine des bases de données classiques [AHV95] ainsi qu'en représentation de connaissances (e.g., par exemple dans les logiques de description [BCM$^+$03] ou dans les graphes conceptuels [CM09]). Lorsqu'un système informatique satisfait l'hypothèse UNA cela signifie que le concepteur du système a pu résoudre des problèmes d'identification d'entités puisqu'il a pu construire une "bijection" entre les identifiants du système et les entités du monde modélisé qui l'intéressent. L'UNA concerne des identifiants dans un système informatique et pas des noms d'entités du monde extérieur. Dans un système, le nom d'une entité n'est souvent qu'un des attributs de l'identifiant représentant cette entité et pas l'identifiant lui-même. De plus, dans de nombreux cas, il n'existe pas de système de nommage d'entités du monde extérieur respectant l'UNA ; en particulier lorsque l'entité nommée est une "entité naturelle" du monde réel (par exemple, une personne, une collectivité, une œuvre...). Le problème est accentué dès que le système a besoin de désigner une grande quantité d'entités nommées et qu'il est partagé par de nombreux utilisateurs.

Dès lors, le problème du choix d'un identifiant pour référencer une entité nommée, aussi appelé *liage*, est un casse-tête qui peut engendrer deux effets : si l'utilisateur est trop sûr de lui, il va utiliser un mauvais identifiant initialement créé pour désigner une autre entité et ainsi introduire des erreurs dans la base ; à l'inverse s'il est trop prudent, il va préférer créer un nouvel identifiant, plutôt qu'utiliser un identifiant existant au risque de se tromper, et ainsi engendrer de l'incomplétude dans la base puisqu'une partie des connaissances concernant une entité sera assertée sur un identifiant et une autre partie sur un autre identifiant.

Dans cet article, nous proposons de ramener ce problème de liage à un problème de vérification des relations de coréférence et différence implicitement assertées dans ces bases. La relation de coréférence dénote deux identifiants qui réfèrent à la même entité et la relation de différence deux identifiants qui réfèrent à deux entités différentes.

Nous étudions ce problème dans le contexte des bases de données documentaires. Ce contexte nous semble intéressant car : il nécessite le référencement de nombreuses entités réelles (des personnes, des collectivités, des œuvres...) ; ces bases intègrent à côté des méta-données de description des documents, des méta-données décrivant des référentiels, i.e. des ensembles typés d'entités nommées importantes (e.g. des personnes, des collectivités, des lieux géographiques, ...) ; ces descriptions sont de "qualité" car produites par des professionnels de l'indexation. Par contre, on observe des erreurs d'identification dans ces bases, se traduisant par l'utilisation dans une description de document d'un identifiant d'une entité nommée ne correspondant pas à l'entité nommée mise en jeu par

BDA septembre 2015, Porquerolles, France.

le document, une erreur de liage, ou par la présence de plusieurs identifiants pour une même entité nommée.

Enfin, la qualité de ces bases est essentielle : elles sont utilisées pour faire du signalement de ressources documentaires, de la bibliométrie, pour calculer des droits pour des ayant-droits, et, plus généralement, pour toute sorte d'interrogation de la base, voire pour établir des liens avec d'autres bases. La qualité de la relation de coréférence/différence est donc fondamentale puisque de nombreux calculs sont basés sur elle. Il faut pouvoir la vérifier et la calculer, par exemple lorsque de nouvelles notices documentaires sont ajoutées.

Notre problème est un cas simple d'un problème fondamental du web sémantique : celui de la vérification de liens *owl :sameAs* et *owl :differentFrom* entre identifiants représentant des entités du monde extérieur. C'est un cas "simple" pour plusieurs raisons : nous nous intéressons à des entités nommées décrites dans des référentiels, les données de base sont des métadonnées décrites dans une unique ontologie qui est un standard international (pas de mappings), seules quelques unes sont en langage naturel, ces métadonnées sont de qualité car construites par des experts, elles sont riches (utilisation de thesaurus) ... mais ce cas "simple" n'est pas pour autant facile à résoudre !

Le premier apport de ce travail est la définition d'un cadre formel pour contrôler la qualité des références aux entités nommées dans un catalogue documentaire. En particulier nous introduisons la notion de référence contextuelle pour désigner la référence à une entité nommée dans le contexte des méta-données d'un document que l'on différencie de la notion de référence d'autorité qui désigne la référence à une entité nommée dans la méta-donnée de description de cette entité. Nous conservons l'identifiant d'origine pour les connaissances relatives à la description de l'entité et créons des identifiants différents pour chacune des références contextuelles. Nous ramenons alors le problème à un problème de contrôle de la relation de coréférence induite par cet éclatement des différentes occurrences d'un identifiant d'entité nommée. D'une part, nous énonçons différentes propriétés que doit satisfaire cette relation. D'autre part nous proposons de la valider en la confrontant à une relation de coréférence construite automatiquement par un système de résolution d'entités modélisant une expertise de liage.

Il existe de nombreux travaux ayant pour objectif de construire une telle relation de coréférence automatiquement que ce soit dans un cadre de fusion de deux bases de données pour détecter des doublons, ou dans le cadre de l'alignement de bases de connaissances ou d'ontologies (en particulier dans le contexte du web sémantique) où il faut repérer que deux identifiants représentent la même entité. La plupart de ces travaux sont basées sur des techniques de classification numérique(cf. [W.E06] ou [ST09]). Dans une telle approche, une entité est décrite par une liste de valeurs d'attributs, souvent considérée comme un vecteur même si l'hypothèse d'indépendance entre attributs n'est pas toujours satisfaite. Les valeurs de ces attributs sont des types de données simples (e.g. des chaines de caractères, des nombres, des dates etc.) pour lesquelles on a des mesures de similarité. A partir de ces mesures de similarité on définit une mesure de similarité entre liste de valeurs d'attributs (une fonction simple des similarités pondérées par l'importance de l'attribut pour le problème d'identification) et une procédure de décision permet de conclure (souvent en utilisant des seuils).

D'autres méthodes sont basées sur des formalisme logique à base de règles qui permettent à partir d'une base de connaissances d'inférer une relation de coréférence/différence (cf. par exemple, [SPR07], [WBGM09], [SPR09], [ARS09], [SPS11]) . Ces méthodes posent le problème de l'acquisition d'un jeu de règles permettant l'inférence d'une telle relation sans introduire trop de couples erronés et sans trop de silence. Des algorithmes d'apprentissage de règles particulières correspondant à des contraintes de clé des bases de données ont été proposés ces dernières années pour aider à l'acquisition de ce type de règles [**?**, **?**, **?**].

Si ces techniques donnent de bons résultats dans le cadre de descriptions riches (i.e. contenant de nombreux attributs) et homogènes, i.e. ce sont les mêmes attributs (à un alignement ontologique prêt) qui sont présents dans deux descriptions comparées, elles ne sont pas directement exploitables dans notre contexte où d'une part les descriptions d'entité nommées sont très succinctes et, d'autre part, on souhaite contrôler la bonne utilisation d'un identifiant d'une entité nommée dans une description de document et pas uniquement détecter des doublons.

La méthode que nous proposons est basée sur la modélisation de l'expertise de professionnels des systèmes documentaires sous la forme de règles d'inférence de prédicats d'identification/différentiation d'identifiants d'entité nommées plus ou moins pertinentes. Ces règles sont basées d'une part sur l'élicitation de caractéristiques importantes (du point de vue de la coréférence) des entités nommées dans les méta-données associées aux documents autant que dans les méta-données associées aux entités nommées, et d'autre part sur la définition de critères d'identification/différentiation de deux identifiants selon quelques unes de ces caractéristiques.

Afin de pallier le peu d'informations contenue dans les descriptions d'entité nommées et leur orthogonalité par rapport à la richesse des informations contenues dans les descriptions de documents, nous enrichissons les descriptions d'entités nommées avec des informations issues des descriptions de documents contenant une référence contextuelle à l'entité dont on est sûr qu'elle est bien coréférente à l'identifiant d'origine. Nous proposons donc d'appliquer un schéma itératif : à partir des prédicats d'identification/différentiation inférées par les règles, nous identifions des relations de coréférences et différences considérées comme sûres ; à partir de ces relations de coréférences sûres nous enrichissons les descriptions d'entité nommées ; et recommençons alors le calcul des prédicats à partir des règles et ainsi de suite jusqu'à stabilité.

Dans la section 2 nous précisons le contexte des catalogues documentaires et les problèmes d'identification d'entités que l'on y trouve. En section 3 nous identifions cinq critères désirés pour un catalogue documentaire permettant d'attester de la qualité de l'identification des entités nommées dans le catalogue et proposons un modèle formel basé sur la vérification/validation de la relation de coréférence. Dans la section 4 nous détaillons le modèle d'expertise utilisé en précisant les notions essentielles (attributs, filtres, critères et règles) et expliquons comment nous l'utilisons pour valider les liens de coréférence. Dans la section 5 nous décrivons le prototype qui a été construit ainsi qu'une première expérimentation.

2. L'IDENTIFICATION D'ENTITÉS DANS LES CATALOGUES DOCUMENTAIRES

Nous présentons dans cette section, quelques spécificités des catalogues documentaires et dressons une typologie des différents types d'erreurs influant sur la qualité de ces bases.

2.1 Catalogue documentaire

Un catalogue documentaire est un système d'information ayant pour objectif de répertorier et retrouver des ressources documentaires. Un tel catalogue est le fruit d'un travail manuel et coopératif d'experts humains (des professionnels de l'indexation) assistés de manière plus ou moins automatique par des outils d'aide à l'indexation. Ces catalogues sont généralement composés de notices dédiées à la description des ressources et de notices dédiées à la des-

cription d'entités nommées utiles à la description des ressources. La figure 1 illustre les différents éléments d'un catalogue documentaire. Les ressources et entités nommées sont des notions externes au système d'information. Ces notices sont exprimées dans des langages et formats ad-hoc (par exemple le format MARC utilisé par les bibliothèques) mais correspondent à des informations semi-structurées que l'on peut représenter dans un langage de représentation des connaissances basé sur la logique du premier ordre (par exemple en RDF).

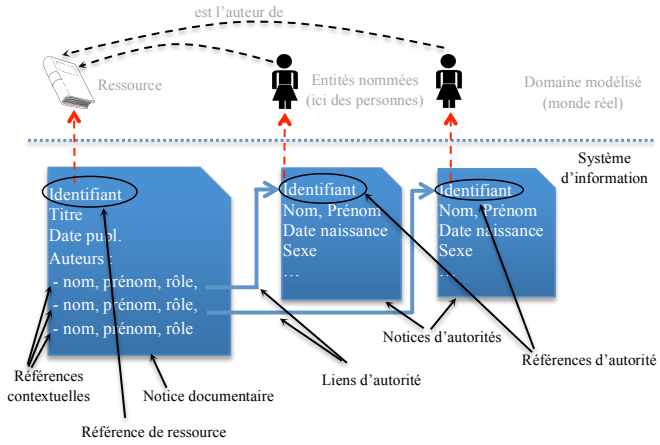

Figure 1: Principaux éléments d'un catalogue documentaire.

2.1.1 *Entité nommée, notice d'autorité, référence d'autorité (RA)*

Une *entité nommée* est une entité importante du monde réel qui est "naturellement" désignée par un *nom* [1]. Les entités nommées peuvent être catégorisées selon leur nature : on parle d'entité nommée personne, collectivité, lieu... Dans la suite de cet article, nous focalisons sur les personnes. Une *référence d'autorité* (RA) est un identifiant qui représente dans un système d'information une entité nommée. Bien qu'il serait souhaitable que deux RA différentes représentent deux entités nommées différentes, cette propriété n'est pas toujours vérifiée dans les catalogues documentaires. Plongées dans un langage logique, les RA sont donc des constantes (ordinaires) sur lesquelles on ne peut pas faire l'hypothèse du nom unique (UNA).

Une *notice d'autorité* est un ensemble d'informations, associés à une RA, exprimées dans un langage de représentation donné et ayant pour vocation de permettre l'identification d'une entité nommée particulière. Une notice d'autorité est en général composée :

- de la RA choisie pour représenter l'entité nommée visée ;
- d'un type permettant d'indiquer la nature de l'entité nommée ;
- des noms avec lesquels l'entité nommée est couramment désignée ;
- d'informations (souvent textuelles) permettant d'identifier l'entité nommée représentée.

1. Un nom est une chaîne de caractères utilisée pour désigner une entité nommée. N'importe quelle chaîne de caractères n'est pas un nom. C'est la forme syntaxique d'une chaîne (par exemple "J. Dupont") ou l'utilisation de cette chaîne dans un contexte particulier (par exemple dans un champ auteur d'une notice documentaire) qui en fait un nom.

2.1.2 *Ressource, notice documentaire, référence de ressource (RR)*

Une *ressource* est une entité qui fait l'objet d'un référencement dans un catalogue documentaire et est directement identifiable, au sens où à partir d'une référence à cette ressource dans le catalogue documentaire, on peut la "restituer" à un utilisateur humain. Selon le cas cette restitution peut être l'envoi de la ressource (cas des ressources électroniques) ou une méthodologie d'accès physique à la ressource (par exemple, localisation précise d'un exemplaire d'un livre dans une bibliothèque). Une *référence de ressource* (RR) est un identifiant qui représente dans un catalogue documentaire une des ressources référencée par le catalogue. Contrairement aux RA, une RR identifie donc précisément la ressource quelle représente puisqu'elle permet l'accès à la ressource. Plongées dans un langage logique, les RR sont des constantes pour lesquelles l'UNA est généralement satisfaite.

Une *notice documentaire* (i.e. une entrée du catalogue) est un ensemble d'informations, associés à une RR, et exprimées dans un langage de représentation donné, ayant pour vocation de décrire – à des fins de signalement – la ressource représentée. Une notice documentaire contient par ailleurs une méthode d'accès à la ressource. Une telle notice est en général composée :

- de la RR choisie pour représenter la ressource visée ;
- d'une méthode d'accès à la ressource ;
- d'un type permettant d'indiquer la nature de la ressource ;
- d'informations sur la ressource permettant de l'identifier sans connaître sa référence : titre, contexte, contributeur, date, sujet... Ces informations utilisent des noms des entités nommées, des RA, des références d'autres ressources, voire des RA ou RR externes au catalogue.

Au sein d'une notice documentaire, les RA permettent d'indiquer les entités nommées utiles à la description de la ressource. Par exemple pour indiquer l'auteur d'un livre, la RA utilisée dans le catalogue pour représenter la personne sera indiquée dans la notice d'autorité. Lors de la construction d'une notice documentaire, cette étape de choix d'une RA, nommée *liage*, est primordiale. Il s'agit pour un indexeur humain (ou parfois un outil de liage automatique) de déterminer, à partir des informations contenues dans le document faisant l'objet d'une notice, parmi les RA du catalogue celle (si elle existe) qui représente l'entité nommée qu'il veut référencer. Les erreurs de liage réalisées lors de cette étape sont à l'origine de la majorité des problèmes de qualité des bases documentaires.

2.1.3 *Lien d'autorité, Référence contextuelle (RC)*

Afin de pouvoir formaliser ces erreurs de liage, nous introduisons la notion de référence contextuelle. Une *référence contextuelle* (RC) est un identifiant qui représente une mention d'une référence à une entité nommée dans une notice documentaire [2]. Plongées dans un langage logique, les références contextuelles sont donc des constantes, sur lesquelles on ne peut pas faire l'hypothèse du nom unique (UNA).

Ainsi un *lien d'autorité*, représentant un liage, peut se voir comme un couple (RC, RA) qui a pour vocation d'identifier l'entité nommée mentionnée par la RC en indiquant la RA présumée représenter cette entité nommée. Dans l'exemple de la figure 1, il y a 3 références contextuelles, mais les 2 premières seulement sont liées. L'ensemble des liens d'autorité d'un catalogue documentaire est donc une fonction partielle des RC dans les RA.

2.2 Typologie des erreurs des catalogues

2. Cet identifiant est construit à partir de la RR identifiant la notice documentaire dans lequel la mention apparaît et de la position de la mention dans la notice.

Les notices prises individuellement ne contiennent généralement pas d'erreurs, dans le sens où les informations qu'elles contiennent sont bien relatives à la ressource ou à l'entité nommée qu'elles représentent. La bonne qualité des informations est due au fait que ces informations ont été saisies manuellement par des professionnels du catalogage. Pour certaines informations textuelles, on peut être confronté à des coquilles mais ce type d'erreurs peut être considéré comme négligeable. C'est l'étape de liage qui engendre la majorité des erreurs. La figure 2 schématise les différents types d'erreurs engendrées par une mauvaise décision de liage.

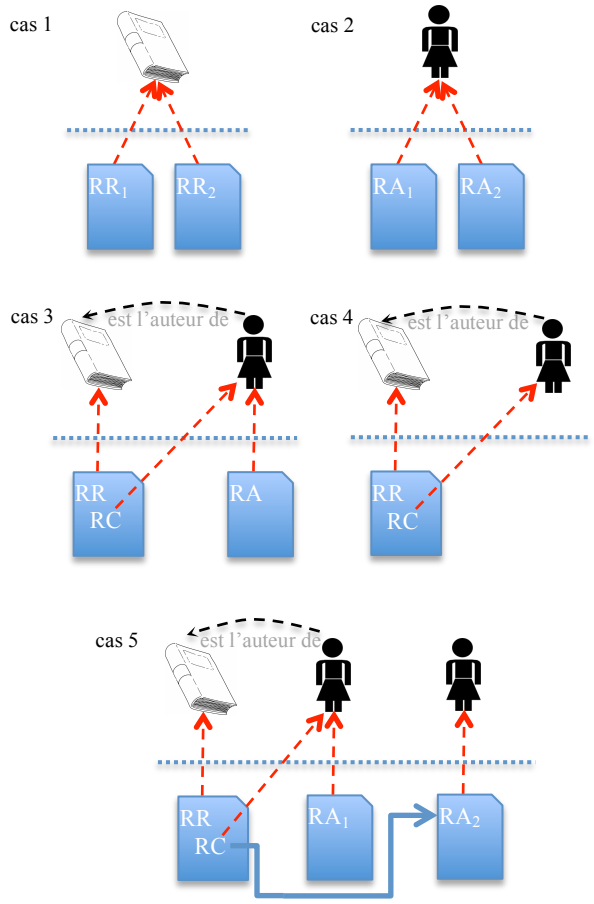

Figure 2: Erreurs référentielles des catalogues documentaires.

Un premier type d'erreur est celui du *doublon de RR* (cf. cas 1 de la figure 2). Il correspond au cas où un indexeur a créé une seconde notice pour un document donné, ou plus souvent à des cas où le catalogue considéré est issu de la fusion de plusieurs catalogues. C'est une erreur peu fréquente car les catalogues intègrent bien souvent des mécanismes de contrôle qui détectent les similitudes entre les informations contenues dans les deux notices. Ce cas d'erreur peut être détecté et réparé aisément en utilisant les techniques classiques de dé-duplication (cf. par exemple, [ARS09], [BG04], [**?**]).

Un second type d'erreur est celui du *doublon de RA* (cf. cas 2 de la figure 2). Il correspond au cas où un indexeur n'a pas identifié formellement une RA existante comme la référence à l'entité nommée qu'il voulait désigner et a créé une nouvelle RA pour cette entité. Il peut aussi correspondre au cas où le catalogue est issu de la fusion de plusieurs catalogues. C'est une erreur plus fréquente et impossible à détecter par les techniques classiques de déduplication car les informations contenues dans les notices d'autorité sont très succinctes (souvent un simple nom et prénom pour une personne). La réparation de cette erreur nécessite donc de définir des mécanismes plus élaborés de détection de doublon et de mettre en œuvre le cas échéant des mécanismes de fusion de notices (cf. par exemple, [HS95], [**?**]).

Un troisième type d'erreur est celui de *l'absence de lien d'autorité* (cf. cas 3 et cas 4 de la figure 2). Il peut correspondre soit au cas où un indexeur n'a pas jugé utile de créer une autorité pour l'entité nommée (car il pensait qu'il n'y aurait pas d'autres mentions de cette entité nommée ou qu'il n'était pas sûr du lien d'autorité à poser et plutôt que de de risquer de créer un doublon a préféré ne pas créer de lien), soit à un cas d'import (de moissonnage) d'un ensemble de notices documentaires sans mise en œuvre d'un processus de liage aux autorités du catalogue. C'est une erreur assez fréquente. On peut distinguer deux sous-cas : le cas où la RA qu'il aurait fallu indiquer est présente dans la base, et celui où elle n'est pas présente dans la base. Dans le premier cas, la réparation passe par la mise en œuvre de mécanismes de liaison automatique. Dans le deuxième cas, cela nécessite en plus que le mécanisme de liaison automatique garantisse qu'aucune des RA de la base ne correspond à la RC et qu'un outil de création automatique de notice d'autorité à partir d'une RC soit disponible.

Un quatrième type d'erreur est celui de *l'erreur de liage* (cf. cas 5 de la figure 2). C'est une erreur très fréquente due d'une part à l'utilisation d'un outil automatique de liage peu performant lors de l'import de catalogue, et d'autre part au fait que des erreurs de liage engendrent d'autres erreurs de liage. En effet, lors de l'étape de liage, les indexeurs observent les différentes notices documentaires liées à une notice d'autorité pour se faire une idée du "genre" de ressources dans lesquelles une RA intervient, cela peut les aider à faire le "bon choix" en raisonnant par analogie avec la ressource de la RC qu'ils cherchent à lier. Dès qu'il existe des erreurs de lien ce raisonnement par analogie devient erroné et tend à engendrer de nouvelles erreurs de liage.

Les conséquences des trois premiers types d'erreurs sont moins importantes que celles du dernier car elles relèvent de l'incomplétude : on peut réparer ces erreurs en complétant le catalogue par des connaissances du type lien "same-as" permettant d'indiquer la coréférence de deux références. Par contre, le quatrième type d'erreur est plus problématique car il introduit une connaissance erronée dans le catalogue : des références se retrouvent coréférentes alors qu'elles ne le sont pas.

Les 4 types d'erreur précédents montrent que l'étape de liage est une étape cruciale dans la construction des catalogues. Le contrôle de la qualité d'un catalogue nécessite donc de contrôler les liens d'autorité et plus généralement de contrôler la relation de coréférence entre les différentes références d'un catalogue. Dans la section suivante, nous proposons un modèle de contrôle de qualité référentielle des catalogues documentaires.

3. QUALITÉ RÉFÉRENTIELLE DES CATALOGUES

Nous considérons que la qualité référentielle d'un catalogue repose sur les cinq critères fondamentaux suivants :

1. **la cohérence locale** : chaque information (en dehors des liens d'autorité) contenue dans une notice est bien relative à l'entité du monde réel décrite par la notice ;
2. **l'absence de doublons de notices documentaires** : le catalogue ne contient pas deux RR représentant la même ressource ;

3. **la complétude du liage** : chaque RC du catalogue est liée à une RA ;
4. **l'absence d'erreurs de liage** : les couples de références (RC,RA) de chaque lien du catalogue représentent bien la même entité nommée ;
5. **l'unicité des autorités** : le catalogue ne contient pas deux RA qui représentent la même entité nommée.

Nous faisons l'hypothèse que les critères 1 et 2 sont toujours vérifiés par les catalogues (les mécanismes d'élaboration des catalogues tendent à éliminer les erreurs relatives à ces critères comme nous l'avons expliqué dans la section précédente) et nous ne considérons dans cet article que les autres critères qui relèvent tous du contrôle de la relation de coréférence (et de son opposée la différence) entre références.

Le critère 3 ne fait appel qu'au contenu du catalogue. Il peut donc être facilement testé et permet de fournir un premier indicateur de qualité relatif à la complétude des liens dans la base : le rapport entre le nombre de RC liées et le nombre de RC.

Les autres critères sont relatifs au "lien de représentation" (i.e. les liens matérialisés en rouge sur les figures 1 et 2) entre RA d'un système d'information et entité nommée du monde réel et ne peuvent donc pas être réellement vérifié. Nous pouvons cependant définir des propriétés formelles du système d'information relatives à la relation de coréférence/différence assertée par les liens d'autorité et utiliser ces propriétés pour construire des indicateurs de qualité référentielle des catalogues.

3.1 Relations d'identification et différentiation

Soit (ID, DI) un couple de relations d'identification/différentiation sur un ensemble $\mathcal{I}$ d'identifiants, nous cherchons à caractériser le fait que de telles relations représente "correctement" les relations de coréférence et différence entre identifiants. Intuitivement, ID est un "bonne relation" d'identification lorsque : si $(i, j) \in ID$ alors i et j représentent avec une "bonne confiance" la même entité, i.e. sont coréférents avec une "bonne confiance". Les guillemets indiquent qu'on ne peut jamais être sûr qu'un couple d'identifiants dits identiques dans un système corresponde à la réalité (i.e. la coréférence) qui est une notion extérieure au système informatique. De même, soit DI une relation de différentiation sur un ensemble d'identifiants, DI est "bonne relation" de différentiation lorsque si $(i, j) \in DI$ alors i et j représentent avec une "bonne confiance" des entités différentes.

On distingue trois types de contrôle de la qualité, relativement à la coréférence et à la différence, d'un couple (ID, DI) : un contrôle structurel, un contrôle logique et un contrôle par rapport à un système "parfait".

Dans la suite nous notons (ID_A, DI_A) les relations binaires duales d'identifiation/différentiation implicitement assertées par les liens d'autorité d'un catalogue. Plus précisément, nous considérons que :

- $\mathcal{I} = \mathcal{C} \cup \mathcal{A}$, i.e. les identifiants $\mathcal{I}$ à considérer sont les ensembles de références d'autorité $\mathcal{A}$ et de références contextuelles $\mathcal{C}$ d'un catalogue ;
- $ID_A = \{(R_1, R_2) \mid (R_1, R_2)\ est\ un\ lien\ d'autorite\}$, i.e. la relation ID_A est constituée de l'ensemble des liens d'autorité ;
- $DI_A = \{(R_1, R_2) \mid R_1 \in \mathcal{A}\ et\ R_2 \in \mathcal{A}\ et\ R_1 \neq R_2\}$, i.e. la relation DI_A est constituée de l'ensemble des couples d'identifiants différents d'autorités.

Dans la section 3.4, nous introduisons un autre couple de relations binaires d'identification/différentiation, construites à l'aide d'une méthode automatique (comme par exemple celle introduite en section 4). Les deux premiers types de contrôle que nous proposons peuvent s'appliquer pour ces deux catégories de couples de relations : dans un cas, pour directement contrôler les liens assertés d'un catalogue ; dans l'autre cas, pour évaluer un système de calcul de relations d'identification/différentiation. Le dernier contrôle proposé met en jeu les deux couples de relations.

3.2 Contrôle structurel

On souhaite qu'un couple (ID, DI) satisfasse le plus possible les contraintes axiomatiques des relations de coréférence (réflexivité, symétrie, transitivité) et de différence (antiréflexivité et symétrie). Nous introduisons dans cette section différentes notions permettant de caractériser la "distance" entre un couple (ID, DI) et un couple de relations d'équivalence et de différence. Cela ne mesure pas le fait que (ID, DI) soit proche de la "vraie" relation de coréférence/différence mais uniquement la proximité avec des relations d'équivalence/différence sur un ensemble donné d'identifiants.

Une première notion, appelée *cohérence structurelle*, est que le couple (ID, DI) soit prolongeable en un couple de relations (ID', DI') de telle sorte que ID' définisse une relation d'équivalence, DI' étant son complémentaire. Nous proposons également une notion moins forte, la *faible cohérence*, qui semble plus réaliste pour qualifier de réels couples de telles relations.

- (ID, DI) est *faiblement cohérent* si $(ID \cap DI) = \emptyset$;
- (ID, DI) est *cohérent* si $(ID^* \cup DI) = \emptyset$, où ID^* désigne la fermeture réflexive, symétrique et transitive de ID ;

Une deuxième notion concerne la *complétude* des relations vis-à-vis de l'ensemble des identifiants considérés. On peut distinguer une notion de *forte complétude*, i.e. chaque couple appartient à ID ou DI, d'une notion plus faible de complétude prenant en compte les contraintes axiomatiques de ces relations, i.e. chaque couple appartient-il à ID^* ou à DI ou son inverse. On peut aussi caractériser la complétude locale des relations ID et DI vis-à-vis des axiomes qu'elles sont censées satisfaire, on dira qu'elles sont *bien définies* :

- (ID, DI) est *fortement complet* si pour tout couple (i, j) d'identifiants de l'ensemble $\mathcal{I}$ considéré on a $(i, j) \in ID \cup DI$;
- (ID, DI) est *complet* si pour tout couple (i, j) d'identifiants de l'ensemble $\mathcal{I}$ considéré on a $(i, j) \in ID^* \cup (DI \cup DI^{-1})$, où DI^{-1} désigne la relation inverse de DI ;
- ID est *bien définie* si $ID = ID^*$, i.e. ID est une relation d'équivalence ;
- DI est *bien définie* si $DI = DI^{-1}$, i.e. DI est symétrique.

Ainsi ces différentes propriétés fournissent des indicateurs de qualité structurelle d'un couple de relations d'identification/différentiation sur un ensemble d'identifiants. Lorsque ces propriétés ne sont pas vérifiées on peut définir des indices numériques d'écart par rapport à ces notions. Par exemple, on peut mesurer l'écart à "bien défini" par : $card(ID^* \setminus ID) + card(DI^{-1} \setminus DI)$; de même, l'écart à "complet" peut être mesuré par le nombre de couples d'identifiants de $\mathcal{I}$ qui n'appartiennent pas à $ID^* \cup DI \cup DI^{-1}$ (ces différentes valeurs pouvant être rendues relatives, par exemple en rapportant l'écart à complet à $card(\mathcal{I})^2$).

3.3 Contrôle logique

Un deuxième niveau de contrôle de ces relations consiste à vérifier la compatibilité des notices (documentaires ou d'autorités) relatives à un couple de références de ID et l'incompatibilité des notices relatives à un couple de références de DI.

Pour cela nous formalisons le contenu d'un catalogue comme une base de connaissances exprimée dans un langage de représentation de connaissances logiquement fondé (que nous supposons

être un sous-ensemble de la logique de premier ordre). Ainsi une notice mentionnant une entité nommée d'identifiant R devient une description, notée D_R déclarant des connaissances sur une entité du monde réel représentée dans le système par la constante logique R (R étant une RA ou une RC selon le contexte).

Nous supposons également que nous disposons de connaissances générales sur le domaine modélisé permettant de définir une notion de cohérence logique d'une description. De nombreuses ontologies relatives aux connaissances représentées dans des catalogues documentaires sont disponibles (Dublin Core, FOAF, FRBR, CRM-CIDOC,...) et peuvent constituer un socle à partir duquel on peut exprimer des contraintes relative à la cohérence d'une (ou plusieurs) descriptions. On peut par exemple exprimer qu'*une personne née à la date d ne peut être l'auteur d'un livre publié à une date d' $\leq$ d*. Nous notons $\mathcal{O}$ de telles connaissances de domaine formalisées dans le langage de représentation de connaissances choisi.

Dans ce cadre, l'hypothèse de vérification du critère 1 de cohérence locale impose que : *tout identifiant R (correspondant à une référence contextuelle ou d'autorité) ayant D_R pour description associée vérifie $\{\mathcal{O}, D_R\}$ est logiquement cohérent* [3].

Avec cette formalisation, on peut définir deux conditions nécessaires (mais non suffisantes) concernant le critère 4 qui reposent sur l'hypothèse de cohérence locale et sur une relation d'identification.

- Condition 1 : *Pour tout couple (R_1, R_2) d'identifiants de ID, on doit avoir $\{\mathcal{O}, D_{R_1}, D_{R_2}, R_1 = R_2\}$ est logiquement cohérent*. Si un couple ne vérifie pas cette propriété, alors R_1 et R_2 ne sont pas coréférents (sous l'hypothèse de cohérence locale).
- Condition 2 : *Pour toute classe d'équivalence $\mathcal{R}$ de ID^*, soit R_0 l'un des identifiants de cette classe d'équivalence, on doit avoir $\bigcup_{R\in\mathcal{R}} \{D_R, R = R_0\} \cup \{\mathcal{O}\}$ est logiquement cohérent*. Si une classe ne vérifie pas cette propriété alors certaines des identifiants de l'ensemble ne sont pas coréférents.

On peut également définir une une condition suffisante (mais non nécessaire) concernant le critère 5 d'unicité de certains identifiants et reposant sur l'hypothèse de cohérence locale :

- Condition 3 (différenciation) : *pour tout couple d'identifiants (R_1, R_2) de DI, on doit avoir $\{\mathcal{O}, D_{R_1}, D_{R_2}, R_1 = R_2\}$ est logiquement incohérent*. Autrement dit, fusionner ces deux identifiants entraînerait une incohérence (ce serait par exemple le cas si l'on identifiait deux autorités personnes de sexe différent alors que l'ontologie déclare l'impossibilité pour une même personne d'être femme et homme).

Ces trois conditions permettent de définir d'autres indicateurs de qualité relatifs à une ontologie donnée. Lorsqu'elles ne sont pas vérifiées, on peut se ramener à des indices numériques. Dans le cadre de (ID_A, DI_A) on peut par exemple calculer : le nombre de liens vérifiant la condition 1 sur le nombre total de liens, le nombre de classes vérifiant la condition 2 sur le nombre de RA, et le nombre de couples de RA vérifiant la condition 3 sur le nombre total de couples de RA différentes.

3.4 Contrôle de la coréférence par confrontation à un modèle d'expertise de liage

Un troisième niveau de contrôle de la coréférence d'un couple (ID, DI) consiste à confronter ce couple à un couple calculé par un système supposé "parfait" ; dans notre cas confronter le couple (ID_A, DI_A) à un couple (ID_S, DI_S) calculé par un système S. Un système "parfait" de calcul de relations d'identification/différentiation, pour un catalogue donné, est un algorithme de calcul d'un couple d'identification/différentiation jugé parfait par des experts sur un échantillon représentatif du catalogue et qu'ils jugeront généralisable à l'ensemble des références du catalogue. Remarquons que pour s'aider à le "juger parfait", les experts pourront utiliser les indicateurs de contrôle structurel et logique précédents.

Pour évaluer la qualité de (ID_A, DI_A) par rapport à (ID_S, DI_S), considéré comme parfait, on peut considérer les trois notions suivantes.

- Une *incomplétude*, qui correspond à une référence $R \in \mathcal{C}$ sans qu'il existe un couple (R, A) dans ID_A alors qu'il existe un couple (R, A) dans ID_S (concerne le critère 3 de complétude du liage).
- Une *erreur de liage*, qui correspond à une référence $R \in \mathcal{C}$ avec (R, A) dans ID_A et soit (R, A) est dans DI_S, i.e. le système parfait dit que R et A ne sont pas coréférentes, ou (R, A) n'est ni dans ID_S ni dans DI_S, i.e. le système parfait dit que R ne peut pas être liée à une autorité (concerne le critère 4 d'absence d'erreurs de liage).
- La *non unicité* des autorités, qui correspond au cas où il existe un couple de références d'autorités (R, R') dans DI_A et dans ID_S ou ni dans ID_S ni dans DI_S (concerne le critère 5).

De la même manière qu'en 3.2 on peut associer des indices numériques à ces différentes notions, les relativiser et les combiner.

Comme expliqué dans l'introduction, les techniques "classiques" de résolution d'entités ne sont pas directement exploitables pour construire de telles relations dans notre contexte où on doit comparer des RA qui ont des descriptions très succinctes à des RC qui ont des descriptions riches. Par ailleurs, nous souhaitons disposer d'une relation calculée de différentiation qui ne soit pas par défaut le complémentaire de la relation d'identification calculée.

Nous présentons dans la section suivante un modèle d'expertise de liage permettant la définition d'un système de calcul de ces deux relations respectant les conditions importantes de cohérence (structurelles et logiques) décrites précédemment.

4. MODÈLE D'EXPERTISE DE LIAGE

Dans cette section nous commençons par décrire rapidement la démarche que nous avons utilisée avec des documentalistes pour représenter les connaissances utiles à l'étude des liages. Ensuite nous décrivons précisément les notions essentielles et les illustrons par des exemples concernant le prototype décrit dans la section 5. Enfin, nous détaillons une stratégie "prudente" permettant de construire un couple de relations d'identification/différentiation (ID_S, DI_S) à partir du modèle d'expertise de liage construit.

4.1 Élicitation des connaissances de liage

Le problème de décision concernant le liage est le suivant : ayant une référence contextuelle X à quelle référence d'autorité Y doit-on la lier ? Il s'agit donc d'exprimer les conditions pour qu'il existe, ou que l'on doive créer, un lien d'autorité $link(X, Y)$.

Une telle connaissance peut s'exprimer naturellement sous la forme d'une règle concluant par $link(X, Y)$ lorsqu'un ensemble d'hypothèses sont satisfaites. Ces règles, qui correspondent à des savoir-faire d'experts, peuvent être plus ou moins pertinentes. Cette plus ou moins forte validité d'une telle règle peut être gérée de différentes manières. On peut, par exemple, pondérer la règle par un coefficient numérique (par exemple une probabilité cf. [**?**] ou [**?**]) ou bien considérer plusieurs prédicats d'identification et de différentiation ordonnés qualitativement (cf. [**?**]). Cette dernière approche permet de rester dans un cadre logique classique mais si, au cours de la mise au point des règles, il semble opportun d'ajouter des prédicats il faut modifier certaines règles ainsi que l'ontologie

3. La cohérence logique d'un ensemble de formules correspond à la classique notion logique d'existence d'une interprétation satisfaisant l'ensemble des formules.

elle même.

Nous avons choisi de considérer comme conclusion un prédicat ternaire $linkID(X, Y, V)$, ayant la sémantique intuitive suivante : les références X et Y réfèrent la même entité avec la certitude V. L'ensemble des valeurs qualitatives de certitude étant muni d'un ordre total (e.g. *always* > *verylikely* > *likely* > *conceivable* > *plausible*). Nous restons ainsi dans un cadre de logique classique tout en introduisant des valeurs qualitatives de certitude et sans avoir à fixer trop tôt le nombre de ces valeurs.

De même pour exprimer la conclusion de règles concernant la différence entre deux références nous avons considéré un prédicat ternaire $linkDI(X, Y, V)$.

A partir des valeurs obtenues avec ces ensembles de règles nous construisons les relations d'identification et de différentiation ID_S et DI_S. Un mécanisme itératif "simple et prudent" pour faire cela est décrit dans la section 4.3. A chaque pas de l'algorithme des liens dits "sûrs" sont construits (i.e. des liens ajoutés à ID_S) et ces liens sont utilisés pour enrichir les informations concernant les références d'autorités de la manière suivante : lorsqu'un lien (X, Y), entre une référence contextuelle X et une référence d'autorité Y est ajouté à ID_S, les informations associées à X sont agrégées à celles déjà présentes dans Y. Cet enrichissement des autorités est un des mécanismes essentiels de notre méthode il est nécessité par le fait que si les informations associées à un document sont assez riches celles associées à une autorité sont plutôt pauvres.

Le corps d'une règle est composé d'un ensemble de conditions. Certaines conditions sont des conditions booléennes simples concernant X ou Y et sont représentées par des prédicats unaires, par exemple *thèse(X)* est satisfaite si et seulement si X est une RC apparaissant dans une thèse. Ces conditions sont appelées des $filtres$. Pour d'autres conditions plus complexes, appelées *critères*, nous avons utilisé la même technique que pour les conclusions des règles, i.e. elles sont représentées par des prédicats dont l'un des arguments est une valeur qualitative, ces valeurs étant munies d'un ordre total. Par exemple : pour exprimer une condition concernant la proximité du nom de deux éditeurs nous utilisons un prédicat ternaire $publisher(X, Y, V)$, les valeurs de V étant ordonnées *very close* > *close* > *neutral* avec un héritage géré implicitement et assurant que si on a $publisher(X, Y, V)$ et $V > V'$ alors on a $publisher(X, Y, V')$. L'approche par prédicats binaires ordonnés qualitativement nous auraient amené à considérer trois prédicats binaires avec les implications suivantes $very\_close\_publisher(X, Y) \rightarrow close\_publisher(X, Y) \rightarrow neutral\_publisher(X, Y)$. Les deux approches sont logiquement équivalentes mais dans une étape de mise au point la première est plus facilement modifiable, par exemple si l'on veut ajouter une quatrième valeur $same$ il suffit d'ajouter $same > very\_close$ dans le mécanisme gérant l'héritage des valeurs.

Pour calculer ces critères nous utilisons des informations du catalogue que nous appelons des $attributs$. En ce qui concerne les références contextuelles ce seront généralement des informations simples obtenues par des requêtes sur la base complétées par des calculs simples. Par exemple, l'attribut $domain$ d'une RC est la liste pondérée et normalisée des domaines dont il est question dans le document de la RC (si 4 domaines sont concernés l'un apparaissant 2 fois et les autres 1 fois, le premier sera muni d'un poids 0.5 et les autres d'un poids 0.25).

Lorsqu'il s'agit d'une référence d'autorité les choses sont plus compliquées. Comme nous l'avons fait remarquer précédemment, les notices d'autorité contiennent généralement peu d'informations. Pour pouvoir comparer des références contextuelles contenant de nombreuses informations avec des références d'autorité succinctes, nous $enrichissons$ les notices d'autorités par des informations issues des RC auxquelles elles sont liées par les liens calculés ID_S. Ainsi, l'attribut $domainSA$ d'une référence d'autorité Y est la liste normalisée des domaines pondérés issue de l'union des valeurs de l'attribut $domain$ des références contextuelles X telles que $(X, Y) \in ID_S$. Si chaque attribut nécessite un type d'agrégation particulier la technique générale consiste à ajouter un item dans un ensemble et à modifier certains poids.

4.2 Formalisation des connaissances

Les connaissances sont représentées en utilisant une base de faits dont le vocabulaire est spécifié par une ontologie de domaine. La base de faits contient l'ensemble des descriptions des références du catalogue. On distingue dans l'ontologie, des connaissances utilisées de manière inférentielles pour compléter les connaissances issues des catalogues documentaires et des connaissances utilisées comme des contraintes (par exemple une personne ne peut pas être une femme et un homme) pour contrôler la cohérence de la base.

4.2.1 Attributs

Un attribut est un "grain de connaissance" qui représente une caractéristique élémentaire intéressante d'un référence contextuelle ou d'une référence d'autorité. Par exemple : la date de publication de la manifestation d'une référence contextuelle, les appellations d'une référence d'autorité... Un attribut a un type (au sens structure de données, par exemple : entier, chaîne de caractères, structure, liste, type énuméré...) et pour une référence donnée un attribut pourra avoir une valeur unique (attribut monovalué) ou une séquence de valeurs de ce type (attribut multivalué). L'attribut *birth* par exemple est un attribut monovalué de type *xsd :date*, tandis que l'attribut *keyword* est un attribut multivalué dont les valeurs sont des *xsd :string*. On se place toujours dans le cadre de l'hypothèse du monde ouvert, aussi le fait qu'une référence n'ait pas une certaine valeur pour un attribut donné ne signifie pas que l'entité référencée n'ait pas la caractéristique représentée par cette valeur.

Différents types d'attributs sont considérés : les attributs natifs dont la valeur est extraite de la base de faits et les attributs calculés dont la valeur est obtenue par agrégation des valeurs d'attributs d'un ensemble de référence.

La valeur d'un attribut natif est extraite par calcul des réponses à une requête conjonctive sur la base de faits saturée par les connaissances inférentielles de l'ontologie. Pour certains attributs, des requête plus complexes du type *If RequêteCondition Then Requête1 Sinon Requête2* mettant en œuvre plusieurs requêtes conjonctives sont nécessaires. Par exemple, la valeur de l'attribut *publicationDate* est définie par trois requêtes, une pour déterminer si le document est une reproduction, si c'est le cas une requête retourne la date de publication de l'original sinon une requête retourne la date de publication du document.

Les attributs calculés permettent l'enrichissement des RA. Ils correspondent à une fonction calculée d'agrégation de valeurs d'attributs issues de plusieurs RC. Initialement, une RA a très peu d'attributs dont les valeurs sont simples à obtenir. Mais à chaque fois qu'un lien sûr est ajouté entre une référence contextuelle X et une référence d'autorité Y il faut modifier certains attributs de Y en les agrégeant avec ceux de X. Par exemple, dans la base documentaire aucun mot clé n'est associé à une notice d'autorité. Nous définissons l'attribut *keywordSA* pour une RA, qui est dit *calculé*, en agrégeant l'ensemble des mots-clés indexant les documents liés par un lien sûr à cette RA, chaque mot-clé étant pondéré par le nombre de RC, relatives à un document contenant ce mot-clé, et liées à cette RA. A chaque fois qu'on ajoute un lien calculé vers RA il faut donc modifier la valeur de son attribut *keywordSA*. La plupart des attributs des RA sont ainsi le résultat d'une opération d'agrégation.

4.2.2 Critères et filtres

Les critères correspondent à des calculs sur les valeurs de certains attributs permettant d'obtenir un indice de proximité ou d'éloignement de deux références en fonction de la valeur de ces attributs. Ces calculs se ramènent souvent à des calculs de similarité entre valeurs d'attributs qui sont ensuite convertis en prédicat qualitatif exprimant la proximité (ou l'éloignement) entre une RC et une RA d'un point de vue particulier correspondant aux attributs comparés. Notons que les attributs calculés des deux références ne sont pas forcément de même nature (mais correspondent cependant à un même point de vue). Par exemple, le critère *pubLifeCriterion* met en jeu les attributs *birth* et *death* d'une RA avec l'attribut *publicationDate* d'une RC.

Les critères sont utilisés dans les hypothèses de règles comme des prédicats qualitatifs exprimant la proximité entre une RC et une RA relativement à certains attributs. Par exemple, le critère *publisherCriterion(X,Y,V)* indiquera que la référence contextuelle X et la référence d'autorité Y sont plus ou moins proches (la valeur de V) en ce qui concerne les éditeurs. Ils seront très proches lorsque l'éditeur de X est l'un des éditeurs de Y (à une variante typographique près), proches lorsque l'éditeur de X a un nom similaire à un des éditeurs de Y, et neutre sinon.

La mise au point de critères pertinents est un travail délicat qui ne se résume pas à la réutilisation de mesures de similarités classiques de l'entité résolution (Levensthein, cosinus, ...). Chaque critère nécessite un travail d'analyse des différentes valeurs des attributs mis en jeu et pour nombre d'entre eux des fonctions de calcul ad-hoc de la valeur du critère ont du être développées. Nous donnons deux exemples de critères dans la section 5.

Les filtres correspondent à des fonctions booléennes de vérification d'une propriété particulière satisfaite par les valeurs de certains attributs d'une référence donnée. Les filtres sont utilisés dans les règles comme des prédicats calculés relatifs à une référence. Par exemple, $literaryGenreFilterRC(X)$ est satisfait si et seulement si le document relatif à X est du genre *literary*.

Notons que si les attributs utiles au calcul de la valeur d'un critère ou d'un filtre ne sont pas renseignés pour une référence donnée, toute règle mettant en jeu ce prédicat ne pourra pas s'appliquer sur cette référence.

4.2.3 Règles $linkID$ et $linkDI$

La partie *décision* de l'expertise des documentalistes dans leur activité de catalogage est représentée par des règles conjonctives et positives (i.e. des règles Datalog) concluant par des prédicats $linkID(X, Y, V)$, indiquant que les références X et Y réfèrent la même entité avec la certitude V ou concluant par $linkDI(X, Y, V)$ indiquant que les références X et Y ne réfèrent pas la même entité avec la certitude V.

Les valeurs de certitude utilisées pour ces deux prédicats, correspondant au troisième argument de ces relations, sont ordonnées de la manière suivante : *always* > *verylikely* > *veryprobable* > *likely* > *probable* > *conceivable* > *plausible*.

Voici deux exemples de règles :

Une règle linkID. Considérons la connaissance experte suivante concernant le liage entre une RC et une RA : *"Si les noms des deux références sont quasi-identiques et si le langage du document est un langage utilisé par l'autorité et si la date de publication du document est compatible avec l'intervalle de vie de l'autorité et si les deux références ont un co-contributeur commun et si le document est de genre littéraire et si l'autorité est auteur de textes littéraires alors il est très probable que RC et RA réfèrent la même personne."*

Cette connaissance est représentée par la règle Datalog suivante : *linkID(X,Y, verylikely) :-rc(X), ra(Y), personNameCriterion(X,Y,veryClose), expLanguageCriterion(X,Y,same), datePubLifeCriterion(X,Y,compatible), cocontribNameCriterion(X,Y,same), literaryGenreFilterRC(X), literaryGenreFilterRA(Y).*

Une règle linkDI. De même, la connaissance : *"Si les noms des deux références ne sont pas quasi-identiques et si la date de publication du document concernant la RC est antérieure à la date de naissance de la RA alors les deux entités sont nécessairement différente"* est représentée par la règle :

linkDI(X,Y,always) :-rc(X),ra(Y), not_personNameCriterion(X,Y,veryClose), datePubLifeCriterion(X,Y,Never).

4.3 Construction des relations ID_S et DI_S

Différentes méthodes de construction des relations calculées ID_S et DI_S peuvent être envisagées à partir du diagnostic posé par les prédicats d'expertise de liage $linkID$ et $linkDI$. Nous décrivons ici une des méthodes simples et prudentse que nous avons utilisée dans le prototype décrit en section 5. Cette méthode est basée sur un algorithme itératif prenant en entrée un ensemble $\mathcal{C}$ de RC et un ensemble $\mathcal{A}$ de RA et calculant une relation ID_S composée de couples (R_1, R_2) où $R_1 \in \mathcal{C}$ et $R_2 \in \mathcal{A}$.

A chaque étape, on cherche à identifier des *liens sûrs*, i.e. des couples (R_1, R_2) à ajouter à ID_S. Un tel lien sûr étant alors considéré comme un lien de coréférence, les connaissances associées à R_1 peuvent être associées à R_2. On "enrichit" donc l'autorité R_2 en lui agrégeant certaines des informations de R_1, techniquement on modifie certains attributs de R_2 en prenant en compte des attributs de R_1. Par ailleurs, R_1 ayant été liée par au moins un lien "sûr", on ne cherche plus à la lier par la suite et on la supprime donc de l'ensemble $\mathcal{C}$. Ainsi, à chaque étape l'ensemble des références contextuelles à lier, $\mathcal{C}$, diminue strictement, alors que l'ensemble des autorités candidates, $\mathcal{A}$, reste lui inchangé. L'itération s'arrête lorsque $\mathcal{C}$ est vide ou lorsque aucun lien sûr nouveau n'a pu été calculé.

Plus précisément, une étape enchaîne les actions suivantes :

1. On applique l'ensemble des règles $linkID$ et $linkDI$ sur l'ensemble des couples de références contextuelles $\mathcal{C} \times \mathcal{A}$.
2. On associe à chaque couple (C_i, A_j) la plus forte valeur $V^{ID}_{i,j}$ de certitude retournée par les règles $linkID$ et $V^{DI}_{i,j}$ la plus forte valeur de certitude retournée par les règles $linkDI$:
   $V^{ID}_{i,j} = max\{V | linkID(C_i, A_j, V)\}$,
   $V^{DI}_{i,j} = max\{V | linkDI(C_i, A_j, V)\}$;
   On associera la valeur *unknown*, considérée comme la plus petite valeur de certitude, lorsqu'aucune valeur de $linkID$ (resp. $linkDI$) n'est obtenue pour un couple.
3. On considère trois valeurs $S_1 > S_2 > S_3$ qui serviront de seuils de certitude de niveau décroissant. Par exemple, S_1= *verylikely*, S_2=*likely* et S_3=*conceivable*. On ajoute un couple (C_i, A_j) à ID_S lorsque :
   - $V^{ID}_{i,j} \geq S_1$, i.e. il est au moins $verylikely$ que C_i et A_j soient coréférentes et
   - $V^{ID}_{i,j} > V^{DI}_{i,j}$, i.e. il y a plus de certitude que C_i et A_j soient coréférentes que différentes et
   - il n'existe pas de RA A_k, avec $A_k \neq A_j$ telle que $V^{ID}_{i,k} > V^{ID}_{i,j}$, i.e. qu'il soit plus pertinent de lier C_i à A_k plutôt qu'à A_j. Notons qu'une RC peut être reliée à plusieurs RA si les valeurs de certitude sont identiques, ce qui devrait dénoter un problème de doublons d'autorités.
4. Pour chaque couple (C_i, A_j) à ID_S, on modifie des attributs de A_j en lui agrégeant ceux de C_i et on supprime C_i de $\mathcal{C}$.

A la fin de l'algorithme, si $\mathcal{C}$ n'est pas vide, on complète ID_S avec les couples (C_i, A_j) tels que :

– $V_{i,j}^{ID} \geq S_2$, i.e. il est au moins $likely$ que C_i et A_j soient coréférentes et
– $V_{i,j}^{DI} \leq S_3$, i.e. il est au plus $conceivable$ que C_i et A_j soient différentes et
– il n'existe pas de RA A_k, avec $A_k \neq A_j$ telle que $V_{i,k}^{ID} > V_{i,j}^{ID}$, i.e. qu'il soit plus pertinent de lier C_i à A_k plutôt qu'à A_j.

Enfin, on complète ID_S en prenant sa fermeture réflexive, symétrique et transitive.

De même, pour construire DI_S, on prend les couples (C_i, A_j) tels que :
– $V_{i,j}^{DI} \geq S_2$ et
– $V_{i,j}^{DI} \leq S_3$.

Et, on symétrise l'ensemble DI_S.

Remarquons que la méthode de construction décrite assure que le couple de relations construites (ID_S, DI_S) est cohérent et que ces relations sont bien définies mais ne garantit pas qu'il est complet (ni, a fortiori, fortement complet).

5. PROTOTYPE

Dans cette section nous décrivons rapidement quelques caractéristiques de l'API que nous avons développée, au-dessus de CoGUI, qui nous a permis de construire le prototype SudocQual dédié au contrôle de la qualité des liens aux autorités de type personne dans le catalogue collectif Sudoc.

5.1 Implémentation du modèle de contrôle

L'implémentation du modèle de contrôle de la qualité référentielle des bases documentaires a été réalisée à l'aide de l'outil CoGUI [4]. CoGUI est un outil interactif doté d'une interface graphique, dédié à la construction de systèmes à bases de connaissances logiquement fondées et à la mise en œuvre de différents raisonnements logiques basés sur une notion fondamentale d'homomorphisme. CoGUI permet la représentation de faits, de hiérarchie de concepts et relations, de règles, de contraintes et de requêtes dans un formalisme de graphes étiquetés permettant en particulier la manipulation de connaissances issues des formalismes du web sémantique (RDF/S, OWL, Sparql...), et fournit des mécanismes généraux d'inférences sur ces primitives. CoGUI embarque par ailleurs un interpréteur de scripts permettant d'intégrer des mécanismes procéduraux au sein des raisonnements logiques. Enfin, CoGUI fournit une API permettant l'adaptation de l'interface à des scénarios dédiés à un problème spécifique de raisonnement.

Une API, baptisée *IFMCR*, permettant la mise en œuvre du modèle expert de décisions de liage décrit dans la section précédente a été implantée au-dessus de l'API CoGUI. IFMCR permet :
– la déclaration d'ensembles de références en fournissant : un identifiant, un type (contextuel ou autorité) et éventuellement la liste des références qui lui sont liées de manière sûre (i.e. des identifiants coréférents) ;
– la déclaration d'attributs natifs en fournissant : un nom d'attribut, une requête indiquant comment récupérer la valeur de cet attribut pour une référence donnée, et éventuellement une fonction de traitement de la valeur permettant de filtrer, nettoyer ou formater les valeurs récupérées ;
– la déclaration d'attributs calculés en fournissant : un nom d'attribut, la liste des attributs que l'on souhaite prendre en compte pour le calcul de cet attribut et une fonction d'aggrégation des valeurs de ces attributs pour les références liées de manière sûre à la référence courante ;
– la déclaration de filtres en fournissant : le nom du filtre, les attributs sur lesquels le filtre est établi et une fonction de satisfiabilité des valeurs de ces attributs par le filtre pour une référence donnée ;
– la déclaration de critères en fournissant : le nom du critère, les attributs sur lesquels le critère est basé et une fonction de calcul d'une valeur de proximité de deux références du point de vue du critère en fonction des valeurs des attributs de ces références ;
– la déclaration de règles de décision d'identification/différentiation composées d'une hypothèse constituée de filtres et valeurs minimales pour un critère et d'une conclusion constituée par un prédicat d'identification ou de différentiation représentant un certain degré de confiance en la décision ;
– la déclaration d'une stratégie de calcul d'un couple de relation d'identification/différentiation basée sur les prédicats d'identification/différentiation (et leur degré de confiance) calculés entre références.

Pour des raisons d'efficacité, IFMCR implémente une application "paresseuse" des règles de décision en ne déclenchant le calcul d'une valeur de critère entre deux références qu'au moment où cette valeur est nécessaire pour vérifier l'hypothèse d'une règle, et en ne déclenchant le calcul d'un attribut qu'au moment où cet attribut est requis pour la vérification d'un filtre ou le calcul d'un critère.

IFMCR s'appuie sur CoGUI car les méta-données d'un catalogue à contrôler ont été représentées comme une base de faits dont la sémantique est précisée par une ontologie de domaine. Les mécanismes de raisonnement basiques de CoGUI assurent la prise en compte des connaissances implicites présentes dans les notices puisque CoGUI permet la prise en compte de règles d'inférences (issues de l'ontologie de domaine) dans les mécanismes d'interrogation et de contrôle de la cohérence de la base de faits.

Ainsi CoGUI fournit à IFMCR les moyens de vérifier les propriétés définies dans la section 3 en utilisant les mécanismes de contrôle de la cohérence d'une base vis-à-vis d'un ensemble de contraintes, de calculer les valeurs des attributs natifs en exploitant son mécanisme d'interrogation de la base, de mettre en œuvre les divers traitements procéduraux en exploitant son interpréteur de scripts, et enfin de charger les méta-données d'un catalogue documentaire dès lors que ces méta-données utilisent les langages standard du web sémantique.

5.2 Mise en œuvre d'un prototype de contrôle de la qualité des liens pour le catalogue du Sudoc

L'approche proposée dans cet article a été mise en œuvre pour le catalogue Sudoc [5], le catalogue du système universitaire français de documentation, géré par l'ABES (agence bibliographique de l'enseignement supérieur) dans le cadre du projet ANR Qualinca [6] dédié à la qualité de l'intégration de catalogues dans de grandes bases documentaires. Le prototype développé en collaboration avec les experts de l'ABES baptisé *SudocQual* se concentre sur le contrôle des liens aux autorités de type personne.

Les méta-données de ce catalogue stockées dans un format UNIMARC ont fait l'objet d'un premier travail de publication au format RDF dans le cadre d'un précédent projet d'enrichissement de notices documentaires externes au Sudoc de références d'autorité du Sudoc [**?**]. Le modèle FRBRoo [7] retenu comme modèle cible de cette publication a fait l'objet d'une formalisation en RDFS que nous avions à cette occasion complétée pour exprimer diverses

4. http ://www.lirmm.fr/cogui/

5. Sudoc : http ://www.sudoc.abes.fr/.
6. Qualinca : http ://www.lirmm.fr/qualinca/
7. FRBRoo : http ://www.cidoc-crm.org/frbr_inro.html

connaissances implicites du modèle (non exprimable en RDFS mais exprimables par des règles et contraintes) et d'autres connaissances spécifiques aux méta-données Sudoc. Ce premier travail de publication a été complété par les experts de l'ABES dans le cadre du nouveau projet Qualinca pour permettre l'identification d'un maximum d'attributs utiles à l'expression du modèle d'expertise de décisions d'identification/différentiation.

Le modèle de décision de relations d'identification/différentiation a été élaboré en collaboration avec les experts de l'ABES. Dans l'état courant, il met en jeu 24 attributs natifs, 18 attributs calculés, 4 filtres, 20 critères, 37 règles d'identification et 14 règles de différentiation.

La définition et la construction des critères a nécessité un travail important et nous donnons deux exemples dans ce qui suit. Certains critères, comme le critère $datePubLifeCriterion$ décrit cidessous, sont assez simples à définir et à construire.

datePubLifeCriterion(X,Y,V). Ce critère exprime la possibilité que la référence contextuelle X soit identique à la référence d'autorité Y relativement à la date de publication de X et aux dates de vie de Y. Ce critère utilise quatre attributs. L'attribut $pubDate$, dont la valeur est la date de publication de la référence contextuelle X, et trois attributs pour la référence d'autorité Y, $birth$ (la date de naissance), on ne dispose parfois que d'une information moins précise $birtUncertain$ qui est un intervalle $[birthInf, birthSup]$ contenant la date de naissance, et parfois d'aucune information sur la date de naissance, et $death$ (la date de décès si elle existe). Intuitivement, si la date de publication est inférieure à la date de naissance le critère aura la valeur $Never$, si elle est contenue dans un sous-intervalle raisonnable de l'intervalle de vie de la RA il aura la valeur $Compatible$, si elle est proche de cet intervalle il aura la valeur $Possible$. Un peu plus précisément, voici le schéma d'algorithme de $datePubLifeCriterion$ que nous avons construit :

SI $(datePubRC = null)$ OU $((birth = null)$ ET $(death = null)$ ET$(birthUncertain = null)))$ RET $NotComparable$;
SI $(birth \neq null$ OU $death \neq null)\{$
SI $(birth = null)$ $dateBirth = dateDeath - 100$ SINON $dateBirth = birth$;
SI $(death = null)$ $dateDeath = dateBirth + 100$ SINON $dateDeath = death$;
SI $(datePubRC < dateBirth + 15)$ RET $Never$;
SINON SI $(datePubRC \geq dateBirth + 15)$ ET $(datePubRC \leq dateDeath)$ RET $Compatible$;
SINON RET $Possible$; $\}$
SINON SI $(birthUncertain \neq null)\{$
SI $(datePubRC < birthInf)$ RET $Never$;
SI $(datePubRC > birthSup + 200)$ RET $Possible$;
RET $NEUTRAL$; $\}$
SINON RET $NotComparable$

Tous les critères ne sont pas aussi simples à définir ou calculer que l'exemple précédent. La similarité entre deux appellations, particulièrement à cause des problèmes posés par des listes de prénoms (prénoms composés, initiales, ordre, ...), est assez difficile à définir même si l'usage de fonctions de similarité standard, comme l'algorithme de Levenshtein pour calculer la proximité de deux chaines de caractères, simplifie sa construction. Nous présentons rapidement un exemple de critère, concernant les domaines importants d'une ressource ou d'un ensemble de ressources, qui nous a conduit à introduire une nouvelle définition de la comparaison de deux ensembles pondérés de termes dépendants, cette dépendance entre termes étant décrite par une fonction de similarité, le calcul s'effectuant ensuite par un algorithme de calcul d'un flot maximal sur un réseau.

domainCriterion La valeur de l'attribut $domain$, pour une RC ou une RA, est un ensemble de couples $D = \{(d_1, p_1), ..., (d_m, p_m)\}$, où les d_i sont des domaines et les p_i des nombres réels entre 0 et 1, exprimant l'importance du domaine pour la référence, vérifiant :

$\forall (t,p) et (t',p') \in D, t \neq t'$, (chaque domaine apparait au plus une fois) et

$$\sum_{\forall (t,p) \in D} p = 1$$

Les domaines (e.g. mathématiques, informatique, électronique, ...), ne sont pas indépendants et leur dépendance est représentée par une mesure de similarité σ, connue, qui vérifie les propriétés suivantes :

- σ est une fonction de $T \times T$ dans l'intervalle des réels $[0, 1]$;
- σ est symétrique : $\forall t, t' \in V \ \ \sigma(t, t') = \sigma(t', t)$;
- $\forall t, t' \in T \ \ \sigma(t, t') = 1$ ssi $t = t'$;

Deux domaines d et d' sont dits *dissimilaires (ou indépendants)* si et seulement si $\sigma(d, d') = 0$. Intuitivement plus $\sigma(d, d')$ est grand plus d et d' sont similaires, et plus il est petit plus ils sont dissimilaires.

Comme les domaines ne sont pas indépendants, une valeur D de l'attribut $domain$ ne peut pas être considérée comme un vecteur et nous avons défini une nouvelle (à notre connaissance) fonction de similarité de la manière suivante : La fonction sim de similarité entre D et D', deux valeurs de l'attribut $domain$, est définie par : $sim(D, D') = max(\sum(x_{ij} \times \sigma(d_i, d_j))$, somme prise pour l'ensemble des couples (d_i, d_j), $d_i \in D$, $d_j \in D'$, et d_i et d_j non dissimilaires, sous les contraintes linéaires suivantes :
pour chaque (d_i, p_i) de D si $d_{i_1}, ..., d_{i_k}$ sont les domaines de D' en correspondance avec d_i on a :
$x_{ii_1} + ...x_{ii_k} \leq p_i$, de même,
pour chaque (d_j, p_j) de D' si $d_{j_1}, ..., t_{j_r}$ sont les domaines de D en correspondance avec d_j on a :
$x_{ji_1} + ...x_{ji_r} \leq p_j$.

Cette fonction a des propriétés souhaitables pour une fonction de similarité d'ensembles de termes pondérés construite à partir d'une fonction de similarité entre termes et nous avons utilisé une implémentation de l'algorithme du simplexe pour la calculer. (pour plus de détails cf. [CLT14]).

Le critère $domainCriterion$ est défini à partir de cette fonction sim en utilisant des seuils de la manière suivante :
IF $(sim \geq 0.6)$ RETURN VeryClose ;
ELSE IF $(sim \geq 0.4)$ RETURN Close ;
ELSE IF $(sim \geq 0.2)$ RETURN Neutral ;
ELSE RETURN Distinct ;

Il n'est pas envisageable d'appliquer directement l'approche proposée sur la base entière qui comporte plus de 10 millions de notices documentaires et plus de 2 millions de notice d'autorité. Comme les autorités personnes ont au minimum un attribut *appellation* dans leur notice qui indique le (ou les) noms sous lesquels ces autorités sont généralement désignées et que par ailleurs toute mention d'une personne dans une notice documentaire indique au minimum cet attribut *appellation* (en le complétant éventuellement d'un lien à une référence d'autorité), nous exploitons cet attribut pour faire un premier partitionnement "grossier" des références de la base. Nous supposons que les appellations permettent de partitionner la base en blocs relatifs à une même appellation (cf. [DSJMB12]). L'ABES ayant développé des services de recherche des autorités personnes relatives à une appellation donnée A (c'est-à-dire dont l'appellation est "proche" de A aux variantes typographiques courantes près : présence d'un ou de plusieurs prénoms, initiales pour les prénoms,

variantes dans les accents, ...) et de recherche des notices documentaires dont l'un des contributeurs à une appellation relative à une appellation A, nous exploitons ces services pour constituer un ensemble $\mathcal{A}$ de références d'autorités et un ensemble $\mathcal{C}$ de références contextuelles susceptibles d'être coréférentes. La base importée est donc réduite aux notices associées à ces références.

Le prototype développé fonctionne en 3 étapes. Tout d'abord un extrait de la base Sudoc relatif à une appellation est exporté en RDF et importé dans le prototype *SudocQual*. Dans une deuxième étape, nous exécutons le modèle d'expertise afin de calculer les relation ID_S et DI_S. Enfin, nous comparons ces relations calculées aux relations ID_A et DI_A assertées dans la base et extrayons nos différents indices de qualité référentielle.

L'exemple suivant illustre le fonctionnement de notre prototype sur l'appellation *"Christian, Bessière"*. 4 RA et 14 RC correspondent à cette appellation dont 1 n'est liée à aucune autorité. La figure 3 montre ID_A, i.e. les liens d'autorités présents dans le catalogue entre ces RC et ces RA. Nous n'avons indiqué que des extraits des attributs. La figure 4 montre le résultat du calcul de la relation ID_S (avant calcul de la fermeture réflexive, symétrique et transitive). L'algorithme itératif s'arrête à la quatrième itération. La relation de différentiation calculée DI_S contient les couples $(C4, A1)$, $(C4, A3)$ et $(C4, A4)$.

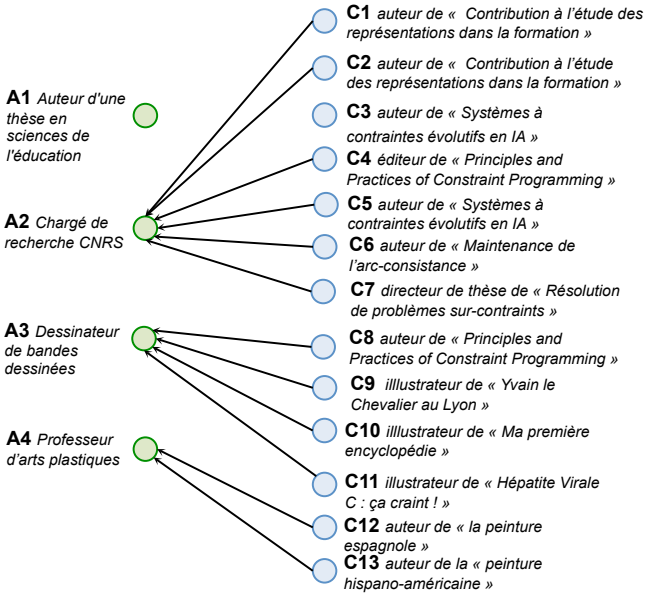

Figure 3: La relation d'identification assertée ID_A pour l'appellation "Christian, Bessière".

Ainsi, les liens d'autorités assertés de $C1$, $C2$ seraient signalés comme erronés et une proposition de remplacerA_2 par $A1$ serait faite ; l'absence de lien pour $C3$ serait signalée et une proposition de poser un lien vers $A2$ faite ; les liens assertés pour $C4$, $C6$ et $C8$ seraient signalés comme non vérifiés. Les liens assertés pour $C5$, $C7$, $C9$, $C10$, $C11$, $C12$, et $C13$ seraient validés.

Sur cet exemple, tous les liens validés sont effectivement corrects. Les liens signalés comme erronés le sont effectivement et les propositions de corrections que la relation ID_S permet de faire pour $C1$, $C2$ et $C3$ sont de bonnes corrections. Les liens non validés de $C4$, $C6$ et $C8$ sont également corrects. Notons également que pour $C4$, l'autorité $A2$ qui est la seule autorité qui n'est pas

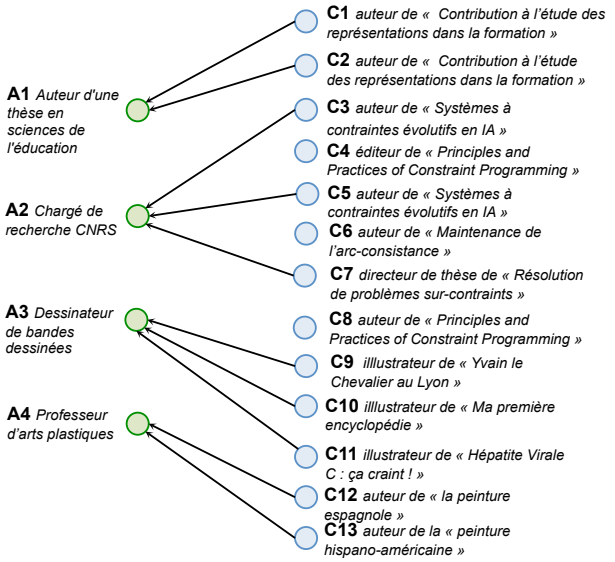

Figure 4: La relation d'identification calculée ID_S (avant fermeture) pour l'appellation "Christian, Bessière".

indiquée comme différente à C_4 dans DI_S est bien coréférente à $C4$.

Les premiers tests sont donc encourageants bien que le modèle d'expertise ne soit pas achevé (il reste en particulier à prendre en compte le calcul de relations d'identification et de différentiation entre deux RC et entre deux RA).

6. CONCLUSION

L'évaluation en cours du prototype SudocQual est basée sur une méthodologie classique en deux grandes étapes. Dans la première étape, des experts de l'ABES ont extrait du catalogue Sudoc un corpus de mise au point dont ils ont corrigé les liens d'autorité et que nous avons utilisé pour construire et évaluer empiriquement avec d'autres experts de l'ABES, attributs, critères et règles. Cette étape est bien avancée et nous a permis de traiter des exemples comme celui décrit dans le paragraphe précédent.

Dans la deuxième étape, qui n'est pas encore commencée, les experts de l'ABES devront choisir et corriger (ce qui prend un temps considérable) un autre corpus jugé pertinent et pour lequel nous disposerons des liens assertés dans le sudoc et des liens corrigés par les experts. Si les résultats de SudocQual sur ce corpus sont très proches (i.e. très bon rappel et très bonne précision) des liens corrigés par les experts, SudocQual pourra être considéré, en mesurant l'écart (rappel et précision) entre les liens assertés et ceux calculés par SudocQual, comme un outil de contrôle de la qualité des liens d'autorité du Sudoc.

Naturellement, notre objectif immédiat est de compléter SudocQual et son évaluation comme indiqué précédemment. Il reste notamment à construire des règles d'identification de deux références contextuelles, des règles d'identification de deux références d'autorité (ceci devrait pouvoir se faire en utilisant exclusivement les attributs et les critères construits pour les liens d'autorité) et à utiliser les résultats de ces règles, par exemple l'existence de doublons dans les notices ou la possible absence d'autorités, pour améliorer la construction des relations (ID_S, DI_S) décrite section 4.3. Il faudra également nous intéresser, toujours dans le cadre du cata-

logue Sudoc, à des liens vers d'autres types d'autorité comme les collectivités.

Ensuite, nous essaierons d'évaluer la généricité de notre approche. En effet, de nombreux chercheurs considèrent qu'il est illusoire d'espérer construire une méthode générale pour les problèmes de coréférence abordés dans cet article. Par exemple Smalheiser et Torvik [ST09] considèrent que : "There is no single paradigmatic author name disambiguation task – each bibliographic database, each digital library, and each collection of publications has its own unique set of problems and issues. For certain purposes (e.g., awarding the Nobel Prize to the author of a breakthrough), it may be very important to achieve a high accuracy of disambiguation. For other purposes (e.g., as an aid to routine information retrieval), it may suffice to assign a high proportion of a person's articles correctly, with little penalty occurring if some articles are missed or mis-assigned."

Nous espérons cependant que notre démarche peut s'appliquer sur d'autres bases documentaires en complétant et modifiant légèrement l'ontologie (qui est basée sur des standard internationaux), en ajoutant des attributs spécifiques et, partie sans doute la plus importante du travail, en adaptant et ajoutant des critères. Mais ceci reste à vérifier !

7. REFERENCES

[AHV95] Serge Abiteboul, Richard Hull, and Victor Vianu. *Foundations of Databases*. Addison-Wesley, 1995.

[ARS09] Arvind Arasu, Christopher Ré, and Dan Suciu. Large-scale deduplication with constraints using dedupalog. In *Proceedings of the 25th International Conference on Data Engineering (ICDE)*, pages 952–963, 2009.

[BCM+03] Franz Baader, Diego Calvanese, Deborah L. McGuinness, Daniele Nardi, and Peter F. Patel-Schneider, editors. *The Description Logic Handbook : Theory, Implementation, and Applications*. Cambridge University Press, 2003.

[BG04] Indrajit Bhattacharya and Lise Getoor. Deduplication and group detection using links. In *ACM SIGKDD Workshop on Link Analysis and Group Detection (LinkKDD)*, 2004.

[CLT14] Michel Chein, Michel Leclère, and Michaël Thomazo. Similarité entre listes de termes dépendants et pondérés. Research report, LIRMM, December 2014.

[CM09] M. Chein and M.-L. Mugnier. *Graph-based Knowledge Representation*. Springer, London, GB, 2009.

[DSJMB12] Anish Das Sarma, Ankur Jain, Ashwin Machanavajjhala, and Philip Bohannon. An automatic blocking mechanism for large-scale de-duplication tasks. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM '12, pages 1055–1064, New York, NY, USA, 2012. ACM.

[HS95] Mauricio A. Hernández and Salvatore J. Stolfo. The merge/purge problem for large databases. *SIGMOD Rec.*, 24(2) :127–138, May 1995.

[SPR07] F. Saïs, N. Pernelle, and M.-C. Rousset. L2r : A logical method for reference reconciliation. In *AAAI*, pages 329–334, 2007.

[SPR09] Fatiha Saïs, Nathalie Pernelle, and Marie-Christine Rousset. Combining a logical and a numerical method for data reconciliation. In Stefano Spaccapietra, editor, *Journal on Data Semantics XII*, volume 5480 of *Lecture Notes in Computer Science*, pages 66–94. Springer Berlin / Heidelberg, 2009.

[SPS11] D. Symeonidou, N. Pernelle, and F. Saïs. Kd2r : a key discovery method for semantic reference reconciliation. In *On the Move to Meaningful Internet Systems : OTM 2011 Workshops*, pages 392–401. Springer, 2011.

[ST09] Neil R. Smalheiser and Vetle I. Torvik. *Annual Review of Information Science and Technology (ARIST)*, volume 43, chapter Author Name Disambiguation. Information Today, Inc, 2009.

[WBGM09] Steven Whang, Omar Benjelloun, and Hector Garcia-Molina. Generic entity resolution with negative rules. *The VLDB Journal*, 18 :1261–1277, 2009. 10.1007/s00778-009-0136-3.

[W.E06] Winkler W.E. Overview of record linkage and current research directions. Technical report, U.S. Census Bureau, 2006.