
Arnaud Giacometti, Béatrice Markhoff, Arnaud Soulet

Découverte de cardinalités maximales significatives dans des bases de connaissances

Volume 3, nº 3-4 (2022), p. 223-251.

<http://roia.centre-mersenne.org/item?id=ROIA_2022__3_3-4_223_0>

**Revue Ouverte d'Intelligence Artificielle**
Volume 3, n° 3-4, 2022, 223-251

# Découverte de cardinalités maximales significatives dans des bases de connaissances

**Arnaud Giacometti[a], Béatrice Markhoff[a], Arnaud Soulet[a]**

[a] Université de Tours - LIFAT, Blois, France
*E-mail :* arnaud.giacometti@univ-tours.fr, beatrice.markhoff@univ-tours.fr, arnaud.soulet@univ-tours.fr.

RÉSUMÉ. — Les bases de connaissances du web sémantique sont générées à partir de plateformes collaboratives ou d'intégration de sources diverses. Cela entraîne évidemment des manques d'information et des erreurs ou incohérences. De plus, dans les programmes d'extraction de connaissances à partir de ces sources il est erroné de considérer que l'absence d'une information dans la base de connaissances équivaut à son inexistence, il faut donc munir la source interrogée d'informations complémentaires permettant de déterminer quand une relation interrogée peut être considérée comme complète. Le volume important de certaines bases nous permet d'utiliser l'inégalité de Hoeffding pour en extraire des règles de cardinalité significatives. Les expérimentations menées sur DBpedia et sur une base de connaissances numismatiques démontrent la faisabilité de l'approche et la pertinence des contraintes extraites.

MOTS-CLÉS. — Découverte de cardinalité, contraintes contextuelles, bases de connaissances.

## 1. INTRODUCTION

Le web des données ouvertes liées, appelé Linked Open Data (LOD), représente une mine d'informations, rassemblées dans des ensembles de données, structurés et de grandes tailles. Leur format (RDF) et leurs descriptions (en RDFS ou en OWL) en font des bases de connaissances ouvertes. Il est crucial pour les applications les interrogeant d'obtenir des réponses significatives par rapport à la réalité. Or, les bases de connaissances du LOD, construites à partir de plateformes collaboratives comme DBpedia [3], ou par des algorithmes d'extraction d'information et d'intégration de sources diverses comme YAGO [32], sont incomplètes et peuvent contenir des faits non valides par rapport à la réalité [34]. Dans [18], l'auteur utilise la notation $W$ pour représenter la réalité qu'une base de données $D$ décrit. Il écrit alors que $D$ est *valide par rapport à* $W$ si $D$ est incluse dans $W$. $D$ est *complète par rapport à* $W$ si $W$ est incluse dans $D$. Dans la lignée de ces travaux, cet article ambitionne de retrouver certaines contraintes sur les données qui sont vraies dans la réalité mais pas forcément dans la ressource considérée, laquelle peut être incomplète et contenir des données incorrectes.

Plus précisément, nos ressources sont des bases de connaissances du LOD et pour en parler nous utilisons les termes usuels en logiques de description [4]. Plutôt qu'une base de données $D$, nous considérons donc une base de connaissances $\mathcal{K} = (\mathcal{T}, \mathcal{A})$ qui décrit une réalité $W$, $\mathcal{T}$ en décrit les *concepts*, leurs relations dénommées *rôles*, ainsi que les règles qui s'y appliquent, tandis que $\mathcal{A}$ en décrit le contenu, c'est-à-dire les *individus* et les assertions afférentes. Afin d'expliciter et de justifier la mesure que nous introduisons pour extraire à partir de $\mathcal{A}$ des contraintes significatives, nous nous référons à $\mathcal{K}^*$, description idéale de $W$, dans le sens où $\mathcal{K}^*$ est valide et complète par rapport à $W$. Nous proposons en effet une méthode pour extraire de $\mathcal{A}$ des règles de $\mathcal{T}^*$ (qui ne sont pas dans $\mathcal{T}$). Il s'agit de *contraintes sur le nombre d'occurrences qu'un rôle peut avoir pour un ensemble de sujets donnés*. En représentation des connaissances, les restrictions numériques précisant le nombre d'occurrences d'un rôle sont connues pour être utiles [5]. Parmi elles, les contraintes de cardinalité maximale permettent de savoir quand toutes les assertions sur un rôle donné pour un individu donné sont présentes dans la base. Cela permet en particulier de compléter les réponses aux requêtes par des informations précises sur leur qualité en termes de *rappel* par rapport à la réalité [23, 29]. Il est illusoire d'espérer des ajouts manuels de telles contraintes d'intégrité dans de grandes bases de connaissances(1), qui soient correctes et suffisantes. Aussi, des techniques de type rétro-ingénierie [26] applicables sur ces grandes bases doivent être considérées, afin de les rechercher systématiquement.

L'extraction de contraintes de cardinalité à partir des données existantes est connue comme un problème important de la rétro-ingéniérie des bases de données relationnelles [26, 33], mais par rapport à ce cadre traditionnel, ce problème est bien plus complexe pour les bases de connaissances du LOD. Pour mieux identifier ces challenges, la table 1.1 fournit des statistiques pour trois rôles de DBpedia, `birthYear` et `parent` dans le contexte du concept `Person`, ainsi que `team` dans le contexte de tout DBpedia que nous notons $\top$, et dans le contexte du concept `FootballMatch`. En particulier, les deux premières colonnes donnent le nombre $n_i$ d'observations d'une cardinalité $i$.

D'abord, ces bases de connaissances contiennent généralement des assertions invalides, que ce soient des assertions fausses ou des assertions dupliquées. De ce fait, la cardinalité maximale observée pour un rôle donné ne saurait être considérée d'emblée comme une cardinalité maximale significative. Par exemple, deux cardinalités maximales significatives sont qu'une personne ait au plus une année de naissance et au plus deux parents. Pourtant dans DBpedia (voir les rôles `birthYear` et `parent` dans la table 1.1), certaines personnes ont jusqu'à 5 années de naissance ou 6 parents ! *Ces quelques assertions invalides ne doivent pas influencer la caractérisation des cardinalités maximales* de ces deux rôles. Ensuite, ces bases de connaissances sont souvent incomplètes. Pour cette raison, la cardinalité la plus observée n'est pas forcément la cardinalité maximale. Typiquement, la plupart des personnes décrites dans DBpedia n'ont qu'un seul parent renseigné (voir le rôle `parent` dans la table 1.1). Toutefois, beaucoup en ont deux et il faut en tenir compte : *la cardinalité maximale d'un rôle ne doit pas être*

(1) [7] présente néanmoins un outil pour le faire sur Wikidata.

TABLE 1.1. Distributions des cardinalités de plusieurs rôles dans DBpedia

`Person` / `birthYear`

| $i$ | $n_i$ | $\tau_i$ | $\widetilde{\tau}_i$ |
|---|---|---|---|
| **1** | **159 841** | **0,999** | **0,996** |
| 2 | 91 | 0,928 | 0,775 |
| 3 | 4 | 0,571 | 0,000 |
| 4 | 2 | 0,667 | 0,000 |
| 5 | 1 | 1,000 | 0,000 |

`Person` / `parent`

| $i$ | $n_i$ | $\tau_i$ | $\widetilde{\tau}_i$ |
|---|---|---|---|
| 1 | 10 643 | 0,529 | 0,518 |
| **2** | **9 392** | **0,991** | **0,975** |
| 3 | 75 | 0,882 | 0,718 |
| 4 | 9 | 0,900 | 0,420 |
| 6 | 1 | 1,000 | 0,000 |

`⊤` / `team`

| $i$ | $n_i$ | $\tau_i$ | $\widetilde{\tau}_i$ |
|---|---|---|---|
| 1 | 1 221 202 | 0,901 | 0,900 |
| 2 | 20 505 | 0,153 | 0,148 |
| 3 | 16 876 | 0,148 | 0,144 |
| ... | ... | ... | ... |
| 20 | 2 | 1,000 | 0,000 |

`FootballMatch` / `team`

| $i$ | $n_i$ | $\tau_i$ | $\widetilde{\tau}_i$ |
|---|---|---|---|
| 1 | 26 | 0,008 | 0,000 |
| **2** | **3 092** | **0,998** | **0,971** |
| 3 | 3 | 0,500 | 0,000 |
| 4 | 2 | 0,667 | 0,000 |
| 5 | 1 | 1,000 | 0,000 |

*sous-estimée au vu de l'ensemble des cardinalités observées*, en l'occurrence la cardinalité maximale du rôle `parent` pour une personne ne doit pas être sous-estimée à 1.

Les effets combinés de données manquantes et de données incorrectes compliquent la détermination de contraintes statistiquement valides. Par exemple, pour résoudre le problème des données incorrectes [20] applique un pré-traitement pour éliminer les observations aberrantes. Cependant les méthodes classiques pour ce faire reposent sur l'hypothèse que les données suivent une distribution normale (ou modérément asymétrique), hypothèse qui n'est pas vérifiée pour la plupart des rôles dans les bases de connaissances considérées et en particulier dans DBpedia comme le montre la table 1.1. Par exemple la cardinalité du rôle `parent` ne suit pas une distribution normale centrée sur 2 du fait de nombreuses données manquantes.

Enfin, il est important de ne pas oublier que les rôles caractérisent des ensembles d'individus précis, formant des contextes particuliers. Par exemple, s'il n'est pas possible de déterminer dans DBpedia une cardinalité maximale pour le rôle `team` en général (contexte `⊤`), il s'avère possible de détecter cette limite pour les matchs de foot (contexte `FootballMatch`) : cette cardinalité maximale correspond alors au nombre d'équipes engagées dans cet événement particulier. *Il est donc essentiel d'identifier les contextes* dans lesquels on puisse détecter des cardinalités maximales. De ce fait, au lieu d'explorer simplement chaque rôle de la base de connaissances, il faut en explorer chaque rôle dans chaque contexte, ce qui donne un espace d'exploration énorme. Il est donc important d'arriver à l'élaguer, sans risquer de perdre des contraintes significatives, mais en évitant de fournir en résultat des contraintes redondantes : si nous détectons que les personnes ont au plus une année de naissance nous ne voulons pas

submerger l'utilisateur de notre système avec d'autres contraintes telles que « les artistes ont au plus une année de naissance », « les scientifiques ont au plus une année de naissance », etc.

Dans cet article, nous répondons aux précédents défis par les trois contributions principales suivantes :

(1) Étant donnée une distribution de cardinalités $(n_i)_{i \geqslant 1}$ observées dans une base de connaissances $\mathcal{K} = (\mathcal{T}, \mathcal{A})$ pour un rôle $R$ dans un contexte $C$, nous proposons d'abord *une méthode de calcul d'une cardinalité maximale significative*. Le premier point de cette contribution est le calcul d'une première estimation de la vraissemblance que la cardinalité $i$ soit maximale dans $\mathcal{K}^*$. Cette estimation, notée $\tau_i$ et appelée taux de cohérence, est calculée en prenant en compte tous les individus de $\mathcal{A}$ pour lesquels le rôle $R$ a une cardinalité *égale ou supérieure* à $i$. Dans $\mathcal{A}^*$, $\tau_i$ est égale à 1 lorsque $i$ est la cardinalité maximale du rôle $R$ dans le contexte $C$, mais ce n'est pas forcément le cas dans $\mathcal{A}$. Le deuxième point de cette contribution est donc une version corrigée de $\tau_i$, *statistiquement valide*, notée $\widetilde{\tau}_i$ et appelée taux de cohérence pessimiste, capable de surmonter les difficultés mises en évidence précédemment concernant les assertions incorrectes et les assertions manquantes dans $\mathcal{A}$. Le calcul de $\widetilde{\tau}_i$ s'appuie sur l'inégalité de Hoeffding, faisant intervenir un niveau de confiance $1 - \delta$. Des exemples d'estimations du taux de cohérence, corrigées ou non, sont indiqués dans les troisième et quatrième colonnes de la table 1.1 pour les rôles `birthYear` et `parent` en considérant le concept `Person` comme contexte, ainsi que pour le rôle `team` en considérant les contextes $\top$ et `FootballMatch`. Précisément, $\tau_i$ est l'estimation fréquentielle de la cohérence et $\widetilde{\tau}_i$ en est la version corrigée s'appuyant sur la borne de Hoeffding. Troisième point de cette contribution, à partir des taux cohérences pessimistes $\widetilde{\tau}_i$, $i \geqslant 1$, nous définissons ce qu'est une cardinalité maximale significative, en utilisant un seuil de cohérence minimum $\min_\tau$.

(2) Nous proposons ensuite C3M[(2)], *un algorithme pour extraire d'une base de connaissances l'ensemble de ses contraintes contextuelles de cardinalité maximale significatives*. Plus précisément, étant donnés une base de connaissances $\mathcal{K}$, un niveau de confiance $1 - \delta$ et un seuil de cohérence minimum $\min_\tau$, cet algorithme explore systématiquement l'ensemble des contraintes contextuelles potentielles de $\mathcal{K}$ et retourne celles qui sont significatives et minimales dans $\mathcal{K}$. Pour cela, nous mettons en évidence deux critères importants qu'une contrainte contextuelle de cardinalité maximale doit vérifier pour être extraite. Ces deux critères sont ensuite utilisés par l'algorithme proposé pour élaguer l'espace de recherche avec des règles sûres (qui ne sont pas des heuristiques).

[(2)] Son implémentation peut être testée ici : `http://c3m.univ-tours.fr/`

(3) Finalement, nous évaluons C3M sur DBpedia et sur COINS, une base de connaissances sur des pièces de monnaie antiques. Non seulement cette évaluation valide C3M et son implémentation mais elle démontre comment l'analyse des contraintes trouvées apporte des informations sur la qualité des données et de leur représentation. Ainsi les résultats expérimentaux montrent que C3M passe à l'échelle de DBpedia et en retourne les contraintes contextuelles de cardinalité maximum significatives : nous estimons que 95 % des contraintes retournées sont correctes par rapport à un ensemble de contraintes sur DBpedia que nous avons annotées manuellement pour nous servir de référence. En plus de mieux caractériser les données, l'analyse des contraintes extraites par C3M permet aussi de mieux comprendre la structure et les choix de conception d'une base de connaissances. Par exemple, des redondances dans les classes ont pu être identifiées dans COINS.

Cet article est une version étendue de [1] avec plusieurs améliorations significatives. L'état de l'art a été complété notamment avec de récentes publications. La partie formelle et l'algorithme ont été modifiés pour bénéficier de la hiérarchie[(3)] des classes de la base de connaissances. Il en résulte un ensemble de contraintes plus compact et sans perte d'information. Enfin, la partie expérimentale a été largement remaniée notamment pour prendre en compte l'analyse de DBpedia.

La suite de cet article est organisée de la façon suivante. Dans la section 2, nous commençons par situer nos propositions par rapport à l'état de l'art. Nous introduisons notre représentation des contraintes contextuelles de cardinalité et nous énonçons formellement notre problème en section 3. Puis, nous expliquons le calcul d'une cardinalité maximale significative en section 4 et nous présentons notre algorithme de découverte de contraintes contextuelles en section 5. Enfin, nous décrivons des expérimentations et analysons leurs résultats en section 6.

## 2. Travaux liés

Cet article s'inscrit dans un vaste mouvement qui vise à augmenter la connaissance sur les données contenues dans les grandes bases de connaissances du web, en termes de validité comme en termes de complétude par rapport à la réalité représentée [23, 25]. Il permet d'enrichir la partie schéma (TBox) de ces bases pour mieux utiliser leur partie données (ABox).

### 2.1. Qualifier la cardinalité des rôles pour un individu

Plusieurs travaux de la littérature visent à qualifier quelle est la cardinalité minimale ou maximale pour un individu donné. Ces travaux consistent donc à enrichir la ABox. On distingue alors les approches endogènes [11, 23] s'appuyant sur les individus de la ABox, et les approches exogènes [17] s'appuyant sur une source extérieure. Les

[(3)] Sachant qu'une classe peut être sous-classe de plusieurs classes, quand nous utilisons dans cet article le terme de hiérarchie, nous entendons le terme de DAG (Direct Acyclic Graph).

auteurs de [11, 23] présentent également des propositions pour déterminer quand un rôle particulier (comme `parent`) manque pour un individu particulier (comme *Obama*). Par exemple, l'hypothèse de complétude partielle stipule que lorsqu'un rôle est renseigné pour un individu, ce rôle est complètement renseigné. Ainsi, comme deux enfants sont renseignés pour Barack Obama dans DBpedia, sous l'hypothèse de la complétude partielle, on considérera qu'il a exactement 2 enfants. Dans [17], une technique de fouille de textes de Wikipedia pour ajouter des précisions sur le degré de complétude des informations dans Wikidata est décrite. Cette approche exogène s'appuie sur des motifs syntaxiques pour relever des propriétés sur les cardinalités. Par exemple, la détection de *The couple's first daughter* pour Barack Obama signifie qu'il faudra au moins une cardinalité minimale supérieure à 1 pour son rôle `hasChild`. Notre proposition est du type endogène car elle traite les données déjà contenues dans les bases de connaissances. Mais surtout, elle ne caractérise pas les rôles par rapport à des *individus* précis mais à des *concepts* définis (au sens des logiques de description). Elle s'avère donc plus générale car les contraintes apprises peuvent être appliquées au niveau des individus de ces concepts (l'inverse n'étant pas vrai).

### 2.2. Qualifier la cardinalité des rôles pour un concept

Dans la littérature, des travaux se sont intéressés à l'enrichissement de la TBox avec de nouvelles assertions ou axiomes permettant de qualifier partiellement ou complètement la cardinalité d'un rôle. En particulier, plusieurs travaux [2, 21, 27, 28] visent l'identification de clés au sens des bases de données relationnelles, en mettant en évidence eux aussi l'importance de la notion de contexte. L'idée est de trouver des axiomes indiquant que tout individu d'un certain concept doit posséder une combinaison de valeurs unique pour une certaine combinaison de rôles. Quand une telle clé est trouvée, elle indique que les rôles impliqués sont renseignés une et une seule fois pour tout individu du contexte et que les combinaison de valeurs sont uniques. Pour ces individus ces rôles ont une cardinalité de 1. Dans [14], les auteurs proposent plus généralement de déterminer automatiquement quels rôles devraient être obligatoirement renseignés pour un concept donné de la base de connaissances. Pour cela ils comparent la densité du rôle pour les individus de ce concept par rapport à sa densité pour les individus d'autres concepts, qui lui sont liés dans la hiérarchie des concepts. Dans le présent article, notre proposition se limite à la détection de contraintes de cardinalité maximale pour un rôle d'un concept, mais contrairement à ces travaux, nous pouvons obtenir des informations sur des cardinalités supérieures à 1 (*e.g.*, une personne a au plus 2 parents).

À notre connaissance, [20] est le seul travail dédié explicitement à la détection de bornes (minimale et maximale) sur la cardinalité des rôles. Cette approche procède en deux temps : élimination des données aberrantes et calcul des bornes. Malheureusement, leur méthode de filtrage repose sur l'hypothèse d'une loi normale sur la distribution des cardinalités (ou faiblement asymétrique) qui n'est que très rarement vérifiée (en particulier, dans les exemples de la table 1.1). À l'inverse, notre approche

bénéficie d'une technique statistique indépendante de toute forme particulière de distribution grâce à l'inégalité d'Hoeffding. Par ailleurs, contrairement à notre approche, ces auteurs ne proposent pas un algorithme pour explorer systématiquement tous les rôles de tous les concepts. Cette exploration est pourtant cruciale et non triviale à cause de l'espace de recherche gigantesque. D'un autre côté, une grande partie de [20] est dédiée à la prise en compte du rôle `owl:sameAs`, qui peut avoir une fonction importante dans certaines bases de connaissances du web. Cela fait partie de nos travaux à venir d'explorer l'impact de ce type de relations génériques entre individus, entre rôles et entre concepts.

### 2.3. Importance de qualifier la cardinalité des rôles

Les différentes sortes d'information supplémentaire sur la qualité des données d'une base de connaissances, en termes de validité comme en termes de complétude par rapport à la réalité représentée, permettent d'améliorer le fonctionnement des applications qui les utilisent, en réduisant le flou de l'hypothèse du monde ouvert. Ainsi c'est pour améliorer la mesure de qualité de règles issues de processus de fouille dans les bases de connaissances du web sémantique que l'*hypothèse de complétude partielle*, déjà évoquée, est définie et utilisée dans [12, 11]. Ces auteurs ont montré que son utilisation rend plus précis le calcul de la confiance associée aux résultats de fouille. Ils ont démontré le besoin, pour la fouille des bases de connaissances du web, de ce qu'ils appellent des *oracles de complétude*, et proposé un certain nombre d'heuristiques pour en définir, comme par exemple la popularité des individus (qui augmente les chances que les faits renseignés sur eux soient complets), etc. Plus récemment, [29] a utilisé des informations sur la cardinalité des rôles pour un individu pour borner la confiance d'une règle.

La fouille de données est loin d'être le seul domaine qui bénéficie de contraintes telles que celles découvertes par notre algorithme. Par exemple, s'appuyant sur des travaux de référence en base de données, les auteurs de [6, 22] et plus récemment [10] proposent de caractériser les réponses obtenues par des requêtes, en fonction des informations connues concernant le degré de complétude de la base de connaissances interrogée, par rapport à la réalité représentée. Notre travail au niveau du schéma permet de dériver de nombreuses informations sur la cardinalité au niveau des individus. La plupart de ces approches pourront donc directement bénéficier des contraintes que nous découvrons.

## 3. Formulation du problème

### 3.1. Bases de connaissances

Dans cet article, nous considérons des *bases de connaissances* $\mathcal{K} = (\mathcal{T}, \mathcal{A})$ où $\mathcal{T}$ et $\mathcal{A}$ sont respectivement les TBox et ABox de $\mathcal{K}$. $\mathcal{T}$ désigne un ensemble d'axiomes terminologiques faisant intervenir des concepts et des rôles, alors que $\mathcal{A}$ désigne l'ensemble des assertions sur les individus. Plus précisément, $\mathcal{A}$ contient des expressions

de la forme $C(a)$ et $R(a, b)$ où $C$ est un *concept*, $R$ est un *rôle*, et $a$, $b$ sont des *individus*.

Dans le cas de la base de connaissances DBpedia, `FootballMatch` et `Person` sont des exemples de concepts et `team` est un exemple de rôle de sa TBox. Par ailleurs, `FootballMatch(1966_FIFA_World_Cup_Final)`[(4)] et `team(1966_FIFA_World_Cup_Final, England_national_football_team)` sont des exemples de faits ou assertions de sa ABox. Le premier indique que `1966_FIFA_World_Cup_Final` est un match de football, alors que le second indique que `England_national_football_team` est une des équipes engagées dans ce match.

Les logiques de description permettent de définir des axiomes pour enrichir la TBox d'une base de connaissances. Par exemple, la relation d'inclusion $\sqsubseteq$ permet d'indiquer qu'un concept $C_1$ est inclus dans un concept $C_2$, noté $C_1 \sqsubseteq C_2$. Plus précisément, une base de connaissances $\mathcal{K}$ implique l'axiome $C_1 \sqsubseteq C_2$ si pour toute interprétation $\mathcal{I}$ de $\mathcal{K}$, $C_1^{\mathcal{I}} \subseteq C_2^{\mathcal{I}}$. Par exemple, les axiomes `Artist` $\sqsubseteq$ `Person` et `Scientist` $\sqsubseteq$ `Person` indiquent respectivement que les concepts `Artist` et `Scientist` sont inclus dans le concept `Person`.

### 3.2. Contraintes contextuelles de cardinalité maximale

Soit $R$ un rôle d'une base de connaissances $\mathcal{K} = (\mathcal{T}, \mathcal{A})$. On considère généralement qu'une contrainte de cardinalité maximale $M$ définie sur $R$ est satisfaite sur $\mathcal{K}$ si pour tout sujet $s$, le nombre d'objets $o$ tels que $R(s, o)$ soit présent dans $\mathcal{K}$ (directement présent dans sa ABox $\mathcal{A}$ ou inférable à partir de ses TBox $\mathcal{T}$ et ABox $\mathcal{A}$) est inférieur ou égal à $M$.

En logiques de description, une telle contrainte peut se représenter par un axiome de la forme $\sqsubseteq$ en utilisant le constructeur de restriction numérique $(\leqslant MR)$. En effet, en terme de logique, une base de connaissances $\mathcal{K}$ implique l'axiome $\exists R.\top \sqsubseteq (\leqslant MR)$ si pour toute interprétation $\mathcal{I}$ de $\mathcal{K}$, $\{x : (\exists y)((x, y) \in R^{\mathcal{I}})\} \subseteq \{x : \#\{y : (x, y) \in R^{\mathcal{I}}\} \leqslant M\}$ où $\#E$ représente la cardinalité d'un ensemble $E$. Il est intéressant de remarquer que $\exists R.\top \sqsubseteq (\leqslant MR)$ si et seulement si $\top \sqsubseteq (\leqslant MR)$, puisque les individus de $\top$ qui ne sont pas sujets de $R$ appartiennent forcément à l'ensemble des individus reliés à moins de $M$ objets par le rôle $R$.

Nous cherchons à identifier des contraintes *contextuelles* de cardinalité maximale, à savoir des contraintes qui ne sont pas nécessairement vérifiées par tous les sujets $s$ d'un rôle $R$, mais par tous les sujets dans un certain contexte, ou plus précisément instances d'un concept, qu'ils soient atomique ou composé, déjà défini dans $\mathcal{K}$ ou pas. Cette notion est introduite formellement dans la définition suivante :

Définition 3.1 (Contrainte contextuelle de cardinalité maximale). — *Étant donnés un rôle $R$, un concept $C$ et un entier $M \geqslant 1$, une contrainte contextuelle de cardinalité maximale définie sur $R$ est une expression $\gamma$ de la forme : $C \sqsubseteq (\leqslant MR)$.*

(4) Pour des raisons de lisibilité nous retirons le préfixe `http://dbpedia.org/ressource/` des URIs des individus de DBpedia.

*Le concept C est appelé le contexte de la contrainte $\gamma$. La contrainte $\gamma$ est satisfaite dans une base de connaissances $\mathcal{K}$ si et seulement si pour toute interprétation $\mathcal{I}$ de $\mathcal{K}$, on a $C^{\mathcal{I}} \subseteq \{x : \#\{y : (x, y) \in R^{\mathcal{I}}\} \leqslant M\}$.*

Par exemple, la contrainte contextuelle `FootballMatch` $\sqsubseteq$ ($\leqslant$ 2 `team`) indique que tous les matchs de foot ont au plus 2 équipes engagées, alors que la contrainte contextuelle `Person` $\sqsubseteq$ ($\leqslant$ 1 `birthYear`) indique que toutes les personnes ont au plus une année de naissance.

Nous cherchons à extraire des contraintes contextuelles de cardinalité maximale qui soient les plus générales possibles. Si on a déjà `Person` $\sqsubseteq$ ($\leqslant$ 1 `birthYear`) il est inutile de chercher si `Artist` $\sqsubseteq$ ($\leqslant$ 1 `birthYear`). De même, si on a déjà `Person` $\sqsubseteq$ ($\leqslant$ 2 `parent`), trouver que `Person` $\sqsubseteq$ ($\leqslant$ 6 `parent`) n'apporte rien non plus.

Définition 3.2 (Contrainte contextuelle minimale). — *Soient deux contraintes contextuelles de cardinalité maximale $\gamma_1 : C_1 \sqsubseteq (\leqslant M_1\ R)$ et $\gamma_2 : C_2 \sqsubseteq (\leqslant M_2\ R)$ définies sur R. La contrainte $\gamma_1$ est dite plus générale que la contrainte $\gamma_2$, noté $\gamma_2 \sqsubset \gamma_1$, ssi $C_2 \sqsubset C_1$[5] et $M_1 \leqslant M_2$, ou bien $C_2 \equiv C_1$ et $M_1 < M_2$. Étant donné un ensemble de contraintes $\Gamma$ définies sur R, une contrainte $\gamma_1 \in \Gamma$ est minimale dans $\Gamma$ s'il n'existe aucune contrainte $\gamma_2$ dans $\Gamma$ plus générale que $\gamma_1$ : $(\nexists \gamma_2 \in \Gamma)(\gamma_1 \sqsubset \gamma_2)$.*

La notion de minimalité a pour objectif de ne pas extraire de contraintes contextuelles qui soient redondantes. Intuitivement, considérons les deux contraintes $\gamma_1$ et $\gamma_2$ introduites dans la définition précédente, et supposons que $\gamma_1$ soit plus générale que $\gamma_2$. Étant donnée une base de connaissances $\mathcal{K}$ dans laquelle les contraintes $\gamma_1$ et $\gamma_2$ sont satisfaites, soit une instance $s$ de $C_2$ dans $\mathcal{K}$. D'après $\gamma_2$, nous savons que pour toute interprétation $\mathcal{I}$ de $\mathcal{K}$, $\#\{o : (s, o) \in R^{\mathcal{I}}\} \leqslant M_2$. Mais comme $\gamma_1$ est plus générale que $\gamma_2$, nous savons par définition que $C_2 \sqsubseteq C_1$. Il en découle que $s$ est aussi une instance de $C_1$ dans $\mathcal{K}$, et d'après $\gamma_1$, que pour toute interprétation $\mathcal{I}$ de $\mathcal{K}$, $\#\{o : (s, o) \in R^{\mathcal{I}}\} \leqslant M_1$, ce qui est une contrainte plus forte que $\#\{o : (s, o) \in R^{\mathcal{I}}\} \leqslant M_2$. En effet, par définition de la minimalité, nous savons que $M_1 \leqslant M_2$. *i.e.* elle ne permet pas de déduire d'information supplémentaire. Dans le cas où les contraintes de cardinalité maximale extraites seraient utilisées pour détecter des inconsistances, la contrainte $\gamma_2$ serait donc inutile et redondante par rapport à la contrainte $\gamma_1$. En effet, toute inconsistance détectée à cause de $\gamma_2$ serait nécessairement détectée à cause de $\gamma_1$. Par ailleurs, il est intéressant de noter que dans les expériences réalisées sur deux bases de connaissances (voir Section 6), il n'est jamais arrivé que soit détectée une contrainte $\gamma_2$ plus générale qu'une contrainte $\gamma_1$. Ces deux remarques justifient de ne rechercher par la suite que des contraintes de cardinalité maximale qui soit minimales.

### 3.3. Problème et défis

Dans l'introduction, nous avons souligné que les bases de connaissances sont souvent incomplètes et contiennent des données erronées. Pour ces raisons, étant

[5] Nous notons $C \sqsubset C'$ pour $C \sqsubseteq C'$ et $C' \not\sqsubseteq C$.

donnée une base de connaissance $\mathcal{K}$, les contraintes de cardinalité satisfaites dans $\mathcal{K}$ne sont pas forcément celles que nous souhaiterions découvrir. Par exemple, en considérant la base de connaissances DBpedia et les statistiques données par la table 1.1, nous ne souhaitons pas découvrir les contraintes (`Person`) $\sqsubseteq$ ($\leqslant$ 6 `birthYear`) et (`Person`) $\sqsubseteq$ ($\leqslant$ 5 `parent`) même si ces dernières sont satisfaites et minimales dans $\mathcal{K}$. Nous voudrions plutôt extraire les contraintes (`Person`) $\sqsubseteq$ ($\leqslant$ 1 `birthYear`) et (`Person`) $\sqsubseteq$ ($\leqslant$ 2 `parent`). Comme indiqué dès l'introduction, nous nous référons à une base de connaissances idéale, dénotée $\mathcal{K}^*$, qui contient tous les faits du monde réel et uniquement les faits du monde réel. Notons qu'en général, nous n'avons pas $\mathcal{K} \subseteq \mathcal{K}^*$ car $\mathcal{K}$ contient des données erronées. Dans ce cadre, notre problème peut être formalisé de la manière suivante :

PROBLÈME 3.3. — *Étant donnée une base de connaissances $\mathcal{K}$, notre objectif est de découvrir l'ensemble de toutes les contraintes contextuelles de cardinalité maximale $C \sqsubseteq (\leqslant M\ R)$ qui sont* minimales *par rapport à la hiérarchie de concepts de $\mathcal{K}$et* satisfaites dans $\mathcal{K}^*$.

Pour résoudre ce problème, nous devons relever les deux défis suivants :

- Le premier défi est de découvrir si une contrainte contextuelle serait vérifiée dans $\mathcal{K}^*$alors que cette base de connaissances idéale est inconnue.
- Le second défi est de parvenir à explorer l'espace de recherche des potentielles contraintes contextuelles de cardinalité maximale dont la taille est gigantesque, en ne conservant que les contraintes qui soient minimales.

## 4. Calcul de contraintes significatives

Pour résoudre le premier défi énoncé dans la section précédente, nous commençons par introduire dans la section 4.1 la notion de taux de cohérence, pour mesurer dans une base de connaissances $\mathcal{K}$la vraisemblance qu'une cardinalité soit maximale dans un contexte donné. Ensuite, ne disposant pas de $\mathcal{K}^*$, nous montrons dans la section 4.2 comment estimer le taux de cohérence d'une contrainte dans $\mathcal{K}^*$par une borne inférieure et ce, uniquement à partir de la base de connaissances $\mathcal{K}$dont on dispose.

### 4.1. Taux de cohérence

Étant donnée une base de connaissances $\mathcal{K} = (\mathcal{T}, \mathcal{A})$, supposons que $i$ soit la cardinalité maximale du rôle $R$ dans le contexte $C$. Soit $s$ un individu de $C$ dans $\mathcal{K}$, complet pour le rôle $R$ dans $\mathcal{K}$ (dans le sens où tous les faits $R(s, o)$ possibles représentant le monde réel sont dans $\mathcal{A}$ ou inférables). Dans le cas où il existe exactement $i$ faits dans $\mathcal{K}$ de la forme $R(s, o)$, cela renforce l'hypothèse que $i$ soit la cardinalité maximale de $R$ dans le contexte $C$. Inversement, s'il existe plus de $i$ faits dans $\mathcal{K}$ de la forme $R(s, o)$, cela affaiblit l'hypothèse que $i$ soit la cardinalité maximale de $R$ dans le contexte $C$. Ainsi dans la table 1.1, pour la classe `Person`, les 85 ($75 + 9 + 1$) individus comportant au moins 3 assertions pour le rôle `parent` affaiblissent l'hypothèse que la cardinalité

maximale soit 2 mais ils restent peu nombreux au regard des 9 392 individus qui ont exactement 2 parents.

En suivant ce raisonnement, nous introduisons la notion de taux de cohérence pour évaluer si une cardinalité $i$ pour le rôle $R$ dans le contexte $C$ a des chances d'être maximale :

Définition 4.1 (Taux de cohérence). — *Étant donnée une base de connaissances $\mathcal{K}$, le taux de cohérence sur $\mathcal{K}$ que la cardinalité i soit maximale pour le rôle R dans le contexte C est le ratio :*

$$\tau_i^{C,R}(\mathcal{K}) = \begin{cases} \frac{n_i^{C,R}}{n_{\geqslant i}^{C,R}} & si\ n_{\geqslant i}^{C,R} > 0 \\ 0 & sinon \end{cases}$$

*où $n_i^{C,R}$ (resp. $n_{\geqslant i}^{C,R}$) représente le nombre de sujets s du contexte C tels que i faits $R(s,o)$ (resp. i faits ou plus) appartiennent à $\mathcal{K}$.*

Par exemple, considérons la contrainte contextuelle `Person` $\sqsubseteq (\leqslant 2$ (*parent*) dans la table 1.1 et calculons le taux de cohérence correspondant $n_{\geqslant 2}^{\texttt{Person,parent}}$. Il est égal à 9 477 (9 477 = 9 392 + 75 + 9 + 1). De cette manière, le taux de cohérence $\tau_2^{\texttt{Person,parent}}$(*DBpedia*) est de 0,991 (*i.e.*, 9392/9477). Par la suite, quand le contexte et la relation sont clairs, nous pouvons les omettre dans les notations. Dans ce cas, $n_i$ et $\tau_i$ désignent respectivement les termes $n_i^{C,R}$ et $\tau_i^{C,R}$ et nous écrirons « taux de cohérence de $i$ » pour parler de $\tau_i$.

Maintenant nous allons formaliser le lien entre le taux de cohérence et la notion de contrainte contextuelle de cardinalité maximale :

Proposition 4.2. — *Étant donnée une hypothétique base de connaissances idéale $\mathcal{K}^* = (\mathcal{T}^*, \mathcal{A}^*)$, le taux de cohérence de la contrainte $C \sqsubseteq (\leqslant M\ R)$ au sein de $\mathcal{K}^*$est égal à 1 si et seulement si $C \sqsubseteq (\leqslant M\ R)$ appartient à $\mathcal{T}^*$, c'est-à-dire que $\tau_M^{C,R}(\mathcal{K}^*) = 1$ si et seulement si $C \sqsubseteq (\leqslant M\ R) \in \mathcal{T}^*$.*

*Démonstration.* — Si $\tau_M^{C,R}(\mathcal{K}^*) = 1$, cela revient à dire que la cardinalité $M$ est atteinte par le rôle $R$ dans le contexte $C$(car $n_{\geqslant M}^{C,R}(\mathcal{K}^*) > 0$) et jamais excédée (car $n_M^{C,R}(\mathcal{K}^*) = n_{\geqslant M}^{C,R}(\mathcal{K}^*)$). La contrainte $C \sqsubseteq (\leqslant M\ R)$ est donc valide dans $\mathcal{A}^*$ et appartient à $\mathcal{T}^*$. Inversement, si $C \sqsubseteq (\leqslant M\ R)$ est dans la TBox $\mathcal{T}^*$, cela veut dire que la règle est vérifiée. La cardinalité $M$ est atteinte par le rôle $R$ dans le contexte $C$(*i.e.*, $n_{\geqslant M}^{C,R}(\mathcal{K}^*) > 0$) et jamais excédée (*i.e.*, $n_M^{C,R}(\mathcal{K}^*) = n_{\geqslant M}^{C,R}(\mathcal{K}^*)$). Par conséquent, $n_M^{C,R}(\mathcal{K}^*)/n_{\geqslant M}^{C,R}(\mathcal{K}^*)$ est exactement égal à 1. □

Cette propriété signifie que les contraintes contextuelles que nous souhaitons découvrir ont un taux de cohérence égal à 1 dans $\mathcal{K}^*$. $\tau_M^{C,R}(\mathcal{K}^*)$ est appelé le taux de cohérence réel. Comme nous ne disposons pas de la base de connaissances idéale $\mathcal{K}^*$, nous pourrions en pratique estimer le taux de cohérence réel avec le taux de cohérence mesuré sur $\mathcal{K}$. Malheureusement, comme pour toute estimation, le taux de cohérence

mesuré dans une base de connaissances est généralement différent du taux de cohérence réel, *i.e.* $\tau_i(\mathcal{K}) \neq \tau_i(\mathcal{K}^*)$. Par exemple, le taux de cohérence $\tau_2^{\texttt{Person,parent}}(\mathcal{K})$ est de 0,991 dans la table 1.1 alors que le taux de cohérence réel d'une cardinalité maximale de 2 pour le rôle `parent` concernant une personne est égal à 1. Plus grave, on a $\tau_6^{\texttt{Person,parent}}(\mathcal{K}) = 1$, alors que le taux de cohérence réel de cette cardinalité est 0 !

### 4.2. Contrainte significative

Intuitivement, si le fait que le taux de cohérence $\tau_6^{\texttt{Person,parent}}(\mathcal{K})$ soit égal à 1 ne fait pas sens, c'est que ce taux est calculé sur un nombre insuffisant d'observations (seulement une personne a 6 parents). De ce fait, l'estimation $\tau_i(\mathcal{K})$ de $\tau_i(\mathcal{K}^*)$ doit être corrigée. Pour ce faire, nous proposons d'utiliser l'inégalité de Hoeffding [13] qui a l'avantage d'être vraie pour toute distribution, et permet de borner l'écart entre une fréquence empirique et une probabilité réelle. Plus formellement, nous introduisons d'abord la définition suivante :

Définition 4.3 (Taux de cohérence pessimiste). — *Étant données une base de connaissances $\mathcal{K}$ et une confiance $1-\delta$, le taux de cohérence pessimiste sur $\mathcal{K}$ que la cardinalité $i$ soit maximale pour le rôle $R$ dans le contexte $C$ est le ratio :*

$$\widetilde{\tau}_i^{C,R}(\mathcal{K}) = \begin{cases} \max\left\{ \frac{n_i^{C,R}}{n_{\geqslant i}^{C,R}} - \sqrt{\frac{\log(1/\delta)}{2n_{\geqslant i}^{C,R}}} \right\} & \text{si } n_{\geqslant i}^{C,R} > 0 \\ 0 & \text{sinon} \end{cases}$$

*où $n_i^{C,R}$ (resp. $n_{\geqslant i}^{C,R}$) représente le nombre de sujets $s$ du contexte $C$ tels que $i$ faits $R(s,o)$ (resp. $i$ faits ou plus) appartiennent à $\mathcal{K}$.*

La propriété 4.4 suivante montre que le taux de cohérence pessimiste est un minorant du taux de cohérence réel (avec une confiance $1-\delta$ donnée) sous deux hypothèses :

- Hypothèse $H1$ : les cardinalités observées dans $\mathcal{K}$ le sont de manière indépendante et identiquement distribuée (*i.i.d.*, independent and identically distributed). Une telle hypothèse est classiquement utilisée en théorie de l'apprentissage [24] pour prendre en compte que les observations dont on dispose sont un échantillon de la réalité.
- Hypothèse $H2$ : la probabilité d'observer une cardinalité $i$ dans $\mathcal{K}$ est égale à la probabilité d'observer une cardinalité $i$ dans $\mathcal{K}^*$. On notera que cette hypothèse n'est pas contredite par le fait que, sur un échantillon de la réalité $\mathcal{K}$ de petite taille, les fréquences d'observation des différentes cardinalités peuvent fortement sous-estimer ou sur-estimer les probabilités réelles d'observation. Qui plus est, un échantillon observé $\mathcal{K}$ étant aléatoire, il y a toujours une probabilité non nulle $\delta$ que cet échantillon soit trompeur, par exemple qu'il ne comprenne que des observations de même cardinalité $m$, alors que $m$ n'est pas la cardinalité maximale réelle.

Proposition 4.4 (Minoration). — *Étant données une base de connaissances $\mathcal{K}$ et une confiance $1-\delta$, sous les hypothèses $H1$ et $H2$, le taux de cohérence réel*

*$\tau_i(\mathcal{K}^*)$ que la cardinalité $i$ soit maximale pour le rôle $R$ dans le contexte $C$ est supérieur au taux de cohérence pessimiste avec une probabilité supérieure à* $(1-\delta)$, i.e. $P(\tau_i(\mathcal{K}^*) \geqslant \widetilde{\tau}_i(\mathcal{K})) \geqslant (1-\delta)$.

*Démonstration.* — En terme de probabilité, si $X$ est une variable aléatoire indiquant pour un sujet $s$ tiré aléatoirement, le nombre de faits $R(s,o)$ appartenant à $\mathcal{K}$, alors $\tau_i$ est une estimation fréquentielle de la probabilité conditionnelle $P(X = i \mid X \geqslant i)$. Tout d'abord, sous les hypothèses $H1$ et $H2$, étant donné un niveau de confiance $1-\delta$, l'inégalité de Hoeffding stipule que $P(\tau_i(\mathcal{K}^*) \in [\tau_i(\mathcal{K}) - \epsilon_i, \tau_i(\mathcal{K}) + \epsilon_i]) \geqslant 1-\delta$ où $\epsilon_i = \sqrt{\frac{\log(1/\delta)}{2n_{\geqslant i}}}$. Par ailleurs, comme le taux de cohérence réel ne peut être que positif, on a : $P(\tau_i(\mathcal{K}^*) \geqslant \widetilde{\tau}_i(\mathcal{K})) = P(\tau_i(\mathcal{K}^*) \geqslant \tau_i(\mathcal{K}) - \epsilon_i)$. Ainsi, on obtient : $P(\tau_i(\mathcal{K}^*) \geqslant \widetilde{\tau}_i(\mathcal{K})) = P(\tau_i(\mathcal{K}^*) \geqslant \tau_i(\mathcal{K}) - \epsilon_i) \geqslant P(\tau_i(\mathcal{K}^*) \in [\tau_i(\mathcal{K}) - \epsilon_i, \tau_i(\mathcal{K}) + \epsilon_i]) \geqslant 1-\delta$, ce qui montre que la propriété est correcte. □

Cette propriété nous munit d'un outil efficace pour approximer le taux de cohérence réel par une borne inférieure (ce qui permettra de prendre des décisions sûres). Par exemple, pour le rôle `parent` de la table 1.1 et un niveau de confiance $1-\delta = 99$ %, nous pouvons noter que le taux de cohérence pessimiste $\widetilde{\tau}_i(\mathcal{K})$ est fortement réduit pour les cardinalités $i = 3$, $i = 4$ et $i = 6$ (en particulier, $\widetilde{\tau}_6^{\texttt{Person,parent}}(\mathcal{K}) = 0$), alors que seul le taux de cohérence pessimiste $\widetilde{\tau}_2^{\texttt{Person,parent}}(\mathcal{K}) = 0,975$ reste proche de 1. Enfin, le fait que le taux de cohérence pessimiste $\widetilde{\tau}_1^{\texttt{Person,parent}}(\mathcal{K}) = 0,518$ reste significativement supérieur 0 (alors que $\tau_1(\mathcal{K}^*) = 0$) tend à montrer que l'hypothèse $H2$ est probablement trop forte (du fait que les bases de connaissances observées sont fortement incomplètes). Toutefois, notre objectif n'est pas de déterminer une bonne estimation des taux de cohérence réels $\tau_i(\mathcal{K}^*)$ pour tout $i \in \mathbb{N}$, mais seulement de détecter s'il existe une cardinalité $i$ telle que $\tau_i(\mathcal{K}^*) = 1$.

Dans ce contexte, étant donné un niveau de confiance $1-\delta$, une fois que l'on dispose pour chaque cardinalité $i$ du taux de cohérence pessimiste $\widetilde{\tau}_i(\mathcal{K})$ (pour un rôle $R$ dans le contexte $C$), nous proposons de considérer qu'une cardinalité $M$ correspond à une cardinalité maximale réelle à deux conditions :

(1) Le taux de cohérence pessimiste $\widetilde{\tau}_M(\mathcal{K})$ est maximal, c'est-à-dire que nous avons $\widetilde{\tau}_M(\mathcal{K}) = \max_{i \geqslant 1} \widetilde{\tau}_i(\mathcal{K})$, et
(2) Le taux de cohérence pessimiste $\widetilde{\tau}_M(\mathcal{K})$ est supérieur à un *seuil minimal de cohérence* $\min_\tau$.

En nous basant sur cette heuristique, nous introduisons finalement la notion de *contrainte significative* :

Définition 4.5 (Contrainte significative). — *Étant donnés un seuil minimal de cohérence* $\min_\tau$ *et un niveau de confiance* $1-\delta$*, une contrainte contextuelle de cardinalité maximale* $C \sqsubseteq (\leqslant M\, R)$ *est dite* significative *connaissant* $\mathcal{K}$ *(ou plus simplement, significative) si et seulement si* $\widetilde{\tau}_M(\mathcal{K}) \geqslant \min_\tau$ *et* $\widetilde{\tau}_M(\mathcal{K}) = \max_{i \geqslant 1} \widetilde{\tau}_i(\mathcal{K})$.

Il est important de noter que pour un rôle $R$ et un contexte $C$, il est possible qu'aucune contrainte significative ne soit déterminée, car pour tout $i$, le taux de cohérence

pessimiste $\widetilde{\tau}_i(\mathcal{K})$ est inférieur au *seuil minimal de cohérence*. Nous verrons dans la section expérimentale (voir Section 6) que ce type de situation est très fréquent, le nombre de contraintes significatives détectées étant en général très inférieur au nombre total de contraintes potentielles.

Quelques exemples d'estimations $\widetilde{\tau}_i$ et de détection de cardinalités maximales contextuelles sont donnés dans la table 1.1. Intuitivement, pour les rôles `birthYear`, `parent` dans le contexte `Person`, on souhaiterait détecter des cardinalités maximales respectives de 1 et 2. Pour un niveau de confiance $1-\delta = 99\,\%$ et un seuil $\min_\tau = 0,97$, on constate que ces cardinalités maximales attendues sont effectivement détectées (cf. lignes en gras dans la table 1.1). De manière intéressante, avec ces mêmes seuils, aucune cardinalité n'est détectée pour `team` dans le contexte le plus général, tandis qu'une cardinalité maximale de 2 apparaît dans le contexte des matchs de football, ce qui là aussi correspond à la réalité.

D'après la définition 4.5, si une contrainte est *significative* connaissant $\mathcal{K}$, cela signifie que son taux de cohérence pessimiste est supérieur à $\min_\tau$ avec une probabilité supérieure à $1-\delta$, et donc qu'elle est très probablement satisfaite dans $\mathcal{K}^*$. Cela nous amène à reformuler le problème 3.3 initialement introduit dans la section 3.3 de la façon suivante :

PROBLÈME 4.6. — *Étant donnée une base de connaissances $\mathcal{K}$, notre objectif est de découvrir l'ensemble de toutes les contraintes contextuelles de cardinalité maximale $C \sqsubseteq (\leqslant M\ R)$ qui sont* minimales *par rapport à la hiérarchie de concepts de $\mathcal{K}$et* significatives *connaissant $\mathcal{K}$.*

## 5. Découverte de contraintes contextuelles de cardinalité maximale

Dans cette section, nous montrons comment résoudre le problème 4.6 défini à la fin de la section précédente. L'espace de recherche étant potentiellement très grand, nous commençons par introduire dans la section 5.1 deux critères d'élagage sûrs permettant de le réduire efficacement. Ensuite, nous présentons dans la section 5.2 un algorithme d'exploration en profondeur utilisant ces deux critères. Finalement, la section 5.3 analyse la complexité de cette approche qui repose sur des requêtes SPARQL.

### 5.1. Critères d'élagage de l'espace de recherche

Dans le cadre proposé, il y a potentiellement un très grand nombre de contraintes contextuelles à considérer, évaluer et comparer. Si $N$ et $P$ sont respectivement les nombres de concepts et de rôles de la base de connaissances considérée, une méthode naïve consisterait à appliquer pour chacune des paires de concepts et rôles existant la méthode de détection d'une contrainte significative présentée à la section précédente. Mais une telle approche nécessiterait $N \times P$ tests, et elle n'est pas faisable pour de très grosses bases de connaissances comme DBpedia (comprenant plus de 483 000 concepts et 60 000 rôles). Néanmoins, il est possible de réduire la taille de cet espace

de recherche en tenant compte des deux critères que doivent satisfaire toute contrainte que nous recherchons, à savoir :

(1) Elle doit être *significative* connaissant $\mathcal{K}$, ce qui veut dire que son taux de cohérence pessimiste doit être supérieur au seuil de cohérence minimum $\min_\tau$. Ceci fournit un premier critère d'élagage formulé par la propriété 5.1.
(2) Elle doit être *minimale* par rapport à la hiérarchie de concepts de $\mathcal{K}$. Ceci fournit un deuxième critère d'élagage formulé par la propriété 5.2.

Selon le premier critère, la contrainte $C \sqsubseteq (\leqslant M\ R)$ doit avoir un taux de cohérence pessimiste supérieur au seuil minimum $\min_\tau$, ce qui ne peut pas être le cas si le contexte $C$ contient trop peu d'individus dans $\mathcal{K}$. En effet, l'intervalle de confiance du taux de cohérence calculé grâce à l'inégalité de Hoeffding est alors très large et sa borne inférieure ne peut être supérieure au seuil $\min_\tau$ imposé. La propriété 5.1 formalise cette intuition :

Proposition 5.1 (Nombre minimal d'observations). — *Soient une base de connaissances $\mathcal{K}$, un seuil minimal de cohérence* $\min_\tau$ *et un niveau de confiance* $(1-\delta)$. *Si on a* $|C \sqcap (\exists R.\top)| < \frac{\log(1/\delta)}{2(1-\min_\tau)^2}$ *pour un concept $C$ et un rôle $R$, alors aucune contrainte de cardinalité maximale $C' \sqsubseteq (\leqslant MR)$ avec $C' \sqsubseteq C$ ne peut être significative connaissant $\mathcal{K}$.*

*Démonstration.* — Démontrons cette propriété par la contraposée en supposant qu'il existe un entier $i$ tel que

$$\widetilde{\tau}_i^{C',R}(\mathcal{K}) = \left(\frac{n_i^{C',R}}{n_{\geqslant i}^{C',R}} - \sqrt{\frac{\log(1/\delta)}{2n_{\geqslant i}^{C',R}}}\right) \geqslant \min_\tau .$$

Cela implique que $1 - \sqrt{\frac{\log(1/\delta)}{2n_{\geqslant i}^{C',R}}} \geqslant \widetilde{\tau}_i^{C',R}(\mathcal{K}) \geqslant \min_\tau$, et donc que $n_{\geqslant i}^{C',R} \geqslant \frac{\log(1/\delta)}{2(1-\min_\tau)^2}$. Par ailleurs, nous avons $|C \sqcap (\exists R.\top)| = n_{\geqslant 0}^{C,R} \geqslant n_{\geqslant i}^{C,R}$ et $n_{\geqslant i}^{C,R} \geqslant n_{\geqslant i}^{C',R}$ (car $C \sqsupseteq C'$). Il en découle que $|C \sqcap (\exists R.\top)| \geqslant n_{\geqslant i}^{C',R} \geqslant \frac{\log(1/\delta)}{2(1-\min_\tau)^2}$, ce qui conclut la preuve par contraposée de la propriété 5.1. □

Cette propriété est particulièrement efficace pour réduire l'exploration de l'espace de recherche. En effet, si pour un contexte $C$ et un rôle $R$ le nombre d'observations $|C \sqcap (\exists R.\top)|$ est insuffisant (il ne satisfait pas le seuil $\log(1/\delta)/(2(1-\min_\tau)^2)$), alors il est impossible de trouver des contraintes $C \sqsubseteq (\leqslant MR)$ significatives connaissant $\mathcal{K}$, mais aussi des contraintes $C' \sqsubseteq (\leqslant MR)$ où $C'$ est un sous-contexte de $C$. En pratique, pour un niveau de confiance de 99 % et un seuil minimal de cohérence de 0, 97, il faut par exemple pour un rôle et un concept donné disposer d'au moins 1 112 observations pour détecter une cardinalité maximale qui soit significative. Ainsi, avec la base de connaissances DBpedia, du fait que seulement 896 faits impliquent des sujets de la classe `Person` pour le rôle `beatifiedDate`, il est certain qu'il est inutile de rechercher une contrainte significative pour le rôle `beatifiedDate` avec `Person` comme contexte, ou tout autre sous-concept de `Person` (tels que `Artist` ou `Scientist`).

Supposons maintenant qu'une contrainte $\gamma$ définie par $C \sqsubseteq (\leqslant M\ R)$ avec $M = 1$ ait été détectée comme significative au cours de l'exploration. Alors, d'après la propriété 5.2 qui suit, il n'est pas nécessaire d'explorer les contraintes $\gamma'$ définies par $C' \sqsubseteq (\leqslant M'\ R)$ où $C'$ est strictement plus spécifique que $C$. Cette propriété découle directement de la définition 3.2 de la minimalité.

Proposition 5.2 (Minimalité des contraintes). — *Soit* $\Gamma$ *un ensemble de contraintes contextuelles de cardinalité maximale. Si l'ensemble* $\Gamma$ *contient une contrainte* $\gamma$ *définie par* $C \sqsubseteq (\leqslant 1\ R)$*, alors aucune contrainte* $\gamma'$ *définie par* $C' \sqsubseteq (\leqslant M'\ R)$ *avec* $C' \sqsubset C$ *ne peut être minimale dans* $\Gamma$.

*Démonstration.* — Cette propriété découle directement des définitions 3.1 et 3.2. En effet, d'après la définition 3.1, nous avons nécessairement $M' \geqslant 1$. Par conséquent, comme nous avons également $C' \sqsubset C$, d'après la définition 3.2, la contrainte $C \sqsubseteq (\leqslant 1\ R)$ est plus générale que la contrainte $C' \sqsubseteq (\leqslant M'\ R)$ avec $C' \sqsubset C$, ce qui implique que la contrainte $C' \sqsubseteq (\leqslant M'\ R)$ ne peut être minimale dans $\Gamma$. □

Cette propriété conduit à un puissant élagage lors de l'exploration des contextes candidats pour les contraintes. En effet, si une contrainte $C \sqsubseteq (\leqslant 1\ R)$ a été détectée comme significative dans $\mathcal{K}^*$, alors il est inutile d'examiner toute les contraintes $C' \sqsubseteq (\leqslant M\ R)$ avec $C' \sqsubset C$. Par exemple, si la contrainte $\texttt{Person} \sqsubseteq (\leqslant 1\ \texttt{birthDate})$ (qui indique que n'importe quelle personne ne peut avoir plus d'une date de naissance) a été détectée, alors il n'est pas nécessaire de rechercher des contraintes plus spécifiques telles que $\texttt{Artist} \sqsubseteq (\leqslant M\ \texttt{birthDate})$ sachant que `Artist` est un sous-concept de `Person`.

### 5.2. C3M : découverte de contraintes de cardinalité maximale

Les critères d'élagage présentés dans la section précédente sont mis en oeuvre dans l'algorithme *C3M*, pour *Contextual Cardinality Constraint Mining*, afin d'explorer efficacement l'espace de recherche des contraintes contextuelles de cardinalité maximale. Il consiste en deux fonctions *C3M-Main* (algorithme 1) et *C3M-Explore* (algorithme 2).

La fonction principale, *C3M-Main*, prend en entrée une base de connaissances $\mathcal{K}$, un niveau de confiance $1 - \delta$ et un seuil minimum de cohérence $\min_\tau$. L'exploration de l'espace de recherche est effectuée indépendamment pour chaque rôle $R$ de la base de connaissances $\mathcal{K}$ (voir la boucle principale de l'algorithme 1 à la ligne 2). Dans une première phase, pour un rôle $R$ de $\mathcal{K}$, l'algorithme 1 conduit une exploration en profondeur de toutes les contraintes contextuelles possibles en considérant tous les contextes (voir la ligne 4 de l'algorithme 1). Cette exploration commence du concept $\top$ de $\mathcal{K}$ [6] en appelant la fonction récursive *C3M-Explore* détaillée ci-après. Comme les concepts de la base de connaissances $\mathcal{K}$ peuvent être subsumés par plusieurs concepts, l'ensemble $\Gamma_R$ des contraintes contextuelles de cardinalité maximale retourné par la

[6] Si la hiérarchie d'une base de connaissances $\mathcal{K}$ comporte plusieurs concepts $\top_i$ sans classe parent, nous considérons qu'un concept plus général $\top$ peut être créé qui subsume tous les concepts $\top_i$, *i.e.* pour tout $i$, on a $\top_i \sqsubseteq \top$.

fonction *C3M-Explore* peut contenir des contraintes qui ne sont pas minimales. Pour cette raison, dans une seconde phase (voir la ligne 6 de l'algorithme 1), la fonction principale *C3M-Main* vérifie pour chaque contrainte $\gamma \in \Gamma_R$ si $\Gamma_R$ contient une contrainte $\gamma'$ qui serait plus générale que $\gamma$. Si ce n'est pas le cas, la contrainte $\gamma$ est vraiment minimale et elle est ajoutée à l'ensemble $\Gamma_m$ des contraintes de cardinalité maximale qui sont minimales. Cet ensemble $\Gamma_m$ est finalement retourné par la fonction *C3M-Main* (voir la ligne 8 de l'algorithme 1).

**Algorithm 1** C3M-Main

**Input:** Une base de connaissances $\mathcal{K}$, un niveau de confiance $1 - \delta$ et un seuil minimal de cohérence $\min_\tau$

**Output:** L'ensemble $\Gamma_m$ de toutes les contraintes contextuelles de cardinalité maximale qui sont significatives et minimales par rapport à $\mathcal{K}$

1: $\Gamma_m := \emptyset$
2: **for all** role in $\mathcal{K}$ **do**
3: *// Explorer en profondeur d'abord la hiérarchie de concepts*
4: $\Gamma_R := C3M\text{-}Explore(\mathcal{K}, R, \top, \infty, \delta, \min_\tau)$
5: *// Ajouter au résultat uniquement les contraintes qui sont minimales*
6: $\Gamma_m := \{\gamma \in \Gamma_R : (\nexists \gamma' \in \Gamma_R)(\gamma \sqsubset \gamma')\} \cup \Gamma_m$
7: **end for**
8: **return** $\Gamma_m$

Nous détaillons maintenant comment la fonction récursive *C3M-Explore* effectue l'exploration en profondeur de l'espace de recherche en bénéficiant des critères d'élagage stipulés par les propriétés 5.1 et 5.2. Pour commencer, la fonction *C3M-Explore* évalue si le nombre d'observations dans $C \sqcap (\exists R.\top)$ est suffisamment important. Si ce n'est pas le cas, nous savons que les contraintes de cardinalité maximale $C' \sqsubseteq (\leqslant M\ R)$ avec $C' \sqsubseteq C$ ne pourront pas être significatives (voir la propriété 5.1) et l'exploration en profondeur s'arrête (ligne 2 de l'algorithme 2). Sinon, le taux de cohérence pessimiste $\widetilde{\tau}_i$ est calculé pour chaque valeur de cardinalité $i$ (lignes 4-6 de l'algorithme 2) et la cardinalité la plus probable $i_M$ est obtenue à la ligne 7. Si le taux de cohérence pessimiste $\widetilde{\tau}_{i_M}$ est inférieur au seuil $\min_\tau$, cela signifie qu'il n'y a pas de contrainte de cardinalité maximale de détectée (à ce niveau de la hiérarchie de $\mathcal{K}$) et $i_M$ est fixé à $\infty$ (ligne 8). Sinon, si $i_M$ est strictement inférieur à $M$ (la cardinalité maximale détectée aux niveaux précédents), cela signifie que nous avons détecté une contrainte de cardinalité maximale $\gamma : C \sqsubseteq (\leqslant i_M\ R)$ qui est potentiellement minimale (comme plusieurs super-concepts peuvent exister dans $\mathcal{K}$, nous devrons vérifier ultérieurement que $\gamma$ est vraiment minimale dans la seconde phase de la fonction *C3M-Main*). Finalement, en utilisant la propriété 5.2, nous savons que si $i_M = 1$, il n'est pas nécessaire de poursuivre l'exploration des descendants du contexte $C$ pour détecter d'autres contraintes de la forme $C' \sqsubseteq (\leqslant M'\ R)$ avec $C' \sqsubseteq C$. Sinon, la fonction *C3M-Explore* est récursivement appelée (ligne 12 de l'algorithme 2) pour explorer l'ensemble des sous-concepts directs de $C$ qui n'ont pas déjà été explorés.

**Algorithm 2** C3M-Explore

**Input:** Une base de connaissances $\mathcal{K}$, un rôle $R$, un contexte $C$, une cardinalité $M$, un niveau de confiance $1-\delta$ et un seuil minimal de cohérence $\min_\tau$

**Output:** Un ensemble $\Gamma$ de contraintes significatives concernant $R$

1: $\alpha := \frac{\log(1/\delta)}{2(1-\min_\tau)^2}$ et $n^{C,R}_{\geqslant 0} := |C \sqcap (\exists R.\top)|$
2: **if** $(n^{C,R}_{\geqslant 0} < \alpha)$ **then return** $\emptyset$
3: $\Gamma := \emptyset$ et $i_{\max} := \arg\max_{i \in \mathbb{N}} \{n^{C,R}_i > 0\}$
4: **for all** $i \in [1..\min\{M, i_{\max}\}]$ **do**
5: $\quad \widetilde{\tau}_i := \max\left\{ \frac{n^{C,R}_i}{n^{C,R}_{\geqslant i}} - \sqrt{\frac{\log(1/\delta)}{2n^{C,R}_{\geqslant i}}}; 0 \right\}$
6: **end for**
7: $i_M := \arg\max_{i \in [1..\min\{M, i_{\max}\}]} \{\widetilde{\tau}_i\}$
8: **if** $(\widetilde{\tau}_{i_M} < \min_\tau)$ **then** $i_M := \infty$
9: **if** $(i_M < M)$ **then** $\Gamma := \{C \sqsubseteq (\leqslant i_M\ R)\}$
10: **if** $(i_M > 1)$ **then**
11: $\quad$ **for all** sous-concept direct $C' \sqsubset C$ non encore exploré **do**
12: $\quad\quad \Gamma := \Gamma \cup$ *C3M-Explore*$(\mathcal{K}, R, C', i_M, \delta, \min_\tau)$
13: $\quad$ **end for**
14: **end if**
15: **return** $\Gamma$

Le théorème suivant indique que l'algorithme *C3M* est correct et complet.

Théorème 5.3. — *Soient une base de connaissances $\mathcal{K}$, un seuil minimal de cohérence* $\min_\tau$ *et un niveau de confiance* $(1-\delta)$. *L'algorithme C3M, composé des fonctions C3M-Main et C3M-Explore, produit l'ensemble de toutes les contraintes de la forme $C \sqsubseteq (\leqslant MR)$ (où $C$ et $R$ sont respectivement des concept et rôle de $\mathcal{K}$) qui sont* significatives *connaissant $\mathcal{K}$ (pour un niveau de confiance $1-\delta$ et un seuil minimal de cohérence* $\min_\tau$*) et* minimales *par rapport à la hiérarchie de concepts définie dans la TBox de $\mathcal{K}$.*

*Démonstration.* — Commençons par montrer que la fonction *C3M-Main* est *correcte*. ll faut montrer que toute contrainte $\gamma^* : C^* \sqsubseteq (\leqslant M^*\ R)$ retournée par la fonction *C3M-Main* est significative connaissant $\mathcal{K}$ et qu'elle est minimale par rapport à la hiérarchie de classes de $\mathcal{K}$. Tout d'abord, il est clair que si une contrainte $\gamma^*$ a été retournée par la fonction fonction *C3M-Main*, alors cela signifie qu'elle a été ajoutée à un moment donné à $\Gamma$ lors d'un appel à la fonction *C3M-Explore*$(\mathcal{K}, R, C^*, M, \delta, \min_\tau)$. De plus, lors de cet appel on a trouvé $i_M = M^* \neq \infty$, ce qui signifie que $\gamma^*$ est significative connaissant $\mathcal{K}$. Supposons maintenant que la contrainte $\gamma^*$ retournée par *C3M-Main* ne soit pas minimale. Si c'est le cas, cela signifie qu'il existe une contrainte $\gamma' : C' \sqsubseteq (\leqslant M'\ R)$ plus générale que $\gamma^*$, *i.e.* telle que $C^* \sqsubset C'$ et $M' \leqslant M^*$, ou $C^* \equiv C'$ et $M' < M^*$, qui soit aussi significative. Notons tout d'abord qu'il est impossible qu'à la fois $\gamma'$ et $\gamma^*$ aient été retournée dans $\Gamma_R$ ( ligne 5 de *C3M-Main*). Sinon, $\gamma^*$ n'aurait pas pu être incluse dans $\Gamma_m$ lors de la recherche des contraintes minimales (ligne 6 de *C3M-Main*). Nous devons maintenant montrer qu'il est impossible que $\gamma^* \in \Gamma_R$ et $\gamma' \notin \Gamma_R$, et allons distinguer deux cas :

(1) La fonction *C3M-Explore*$(\mathcal{K}, R, C', M, \delta, \min_\tau)$ a été appelée, mais la contrainte $\gamma'$ n'a pas été ajouté à $\Gamma$ (ligne 9) (puisque $\gamma'$ n'a pas été retournée dans $\Gamma_R$). Dans ce cas, lors de cet appel, on a trouvé $i_M = M' \neq \infty$ (car $\gamma'$ est supposée significative), mais $i_M \geqslant M$, ce qui implique que $M \neq \infty$. Il en découle que lorsque la fonction *C3M-Explore*$(\mathcal{K}, R, C', M, \delta, \min_\tau)$ a été appelée, avait été nécessairement explorée précédemment et ajoutée à $\Gamma$ une contrainte $\gamma'' : C'' \sqsubseteq (\leqslant M''R)$ avec $C' \sqsubset C'' \sqsubseteq \top$ et $M'' = i_M \neq \infty$. Mais dans ce cas, la contrainte $\gamma''$ aurait été retournée dans $\Gamma_R$, et $\gamma^*$ moins générale que $\gamma''$ n'aurait pas pu être ajoutée à $\Gamma_m$, ce qui contredit l'hypothèse de départ.

(2) La fonction *C3M-Explore*$(\mathcal{K}, R, C', M, \delta, \min_\tau)$ n'a pas été appelée. Cela ne peut être lié au fait que pour tout antécédent de $C'$ dans la hiérarchie de $\mathcal{K}$ aucune contrainte significative ait été détectée (sinon $\gamma'$ ne pourrait être significative d'après la propriété 5.1). Donc, si *C3M-Explore*$(\mathcal{K}, R, C', M, \delta, \min_\tau)$ n'a pas été appelé, c'est qu'il existe une contrainte $\gamma'' : C'' \sqsubseteq (\leqslant 1R)$ avec $C^* \sqsubset C' \sqsubset C''$ qui a été détectée précédemment comme significative et qui a été ajoutée à $\Gamma$, donc retournée dans $\Gamma_R$. Mais alors, la contrainte $\gamma^*$ moins générale que $\gamma''$ n'aurait pas pu être retournée par *C3M-Main* (car non conservée lors de la recherche des contraintes minimales de $\Gamma_R$), ce qui contredit l'hypothèse que la contrainte $\gamma^*$ retournée par *C3M-Main* ne soit pas minimale et finit de prouver que l'algorithme *C3M-Main* est correct.

Montrons maintenant que l'algorithme *C3M-Main* est *complet*, *i.e.* qu'il retourne toutes les contraintes $C \sqsubseteq (\leqslant MR)$ significatives et minimales. Supposons qu'une telle contrainte $\gamma^* : C^* \sqsubseteq (\leqslant M^*R)$ n'ait pas été retournée par l'algorithme *C3M-Main*. Tout d'abord, si $\gamma^*$ avait été retournée par l'appel de la fonction *C3M-Explore* (ligne 5), elle aurait nécessairement été incluse dans $\Gamma_m$ lors de la recherche des contraintes minimales (ligne 6 de *C3M-Main*). Sinon, cela contredirait sa minimalité. Par conséquent, la contrainte $\gamma^*$ n'a pas pu être retournée lors de l'appel de la fonction *C3M-Explore* (ligne 5). Faisons maintenant cette hypothèse. D'après les propriétés 5.1 et 5.2, il est tout d'abord clair que la contrainte $\gamma^*$ a nécessairement été explorée, *i.e.* la fonction *C3M-Explore*$(\mathcal{K}, R, C^*, M, \delta, \min_\tau)$ a été appelée. Sinon, soit elle ne pourrait pas être significative, soit il existerait une contrainte $\gamma : C \sqsubseteq (\leqslant 1R)$ significative avec $C^* \sqsubset C$, ce qui contredirait à nouveau la minimalité de $\gamma^*$. Maintenant, si la fonction *C3M-Explore*$(\mathcal{K}, R, C^*, M, \delta, \min_\tau)$ a été appelée, alors que la contrainte $\gamma^*$ n'a pas été ajoutée à $\Gamma$ (ligne 9 de *C3M-Explore*), cela signifie que $i_M = M^* \geqslant M$ et donc que $M \neq \infty$ (car $i_M = M^* \neq \infty$, $\gamma^*$ étant supposée significative). Il en découle que lorsque la fonction *C3M-Explore*$(\mathcal{K}, R, C^*, exp, M, \delta, \min_\tau)$ a été appelée, avait été explorée précédemment une contrainte $\gamma : C \sqsubseteq (\leqslant MR)$ où $C^*$ est un descendant direct de $C$ et $i_M = M \neq \infty$. Cette contrainte $\gamma$ est significative et plus générale que $\gamma^*$, ce qui contredirait à nouveau la minimalité de $\gamma^*$, et termine la preuve de la complétude de l'algorithme *C3M-Main*.

□

### 5.3. Stratégie d'implémentation et complexité

Dans notre implémentation de la fonction C3M-Explore, nous avons appliqué une approche client-serveur où les distributions de cardinalité $n_i^{C,R}$ sont calculées par interrogation en SPARQL d'une base de connaissances localisée sur un serveur. Par exemple, la requête ci-dessous calcule la distribution des cardinalités de la relation `birthYear` pour les individus de la classe `Person` dans DBpedia :

```
SELECT ?cardinality (COUNT(?entity) AS ?count)
FROM <http://dbpedia.org>
WHERE {
  SELECT ?entity (COUNT(?relation) AS ?cardinality)
  WHERE {
    FILTER EXISTS {?entity rdf:type dbo:Person} .
    ?entity dbo:birthYear ?relation
  }
  GROUP BY ?entity
}
GROUP BY ?cardinality
ORDER BY DESC(?cardinality)
```

Dans un tel cadre, la complexité de la fonction C3M-Explore en nombre de requêtes sur le serveur est en $O(|\mathcal{H}|)$ où $|\mathcal{H}|$ représente le nombre de concepts dans la hiérarchie de concepts $\mathcal{H}$ de la TBox de $\mathcal{K}$. Dans le pire des cas, côté client, la complexité en nombre d'opérations est en $O(|\mathcal{H}| \times i_{\max})$ où $i_{\max}$ représente l'entier maximal pour lequel il existe au moins un sujet $s$ tel que $i_{\max}$ faits $R(s, o)$ appartiennent à la base de connaissances $\mathcal{K}$, *i.e.* $i_{\max} = \arg\max_{i \in \mathbb{N}} \{n_i^{\top,R} > 0\}$.

## 6. Expérimentations

Nous présentons nos expérimentations sur deux bases de connaissances aux caractéristiques bien différentes. La première est DBpedia, qui nous permet de montrer que nos propositions passent à l'échelle des grandes ressources publiques du web sémantique et peuvent donc contribuer à rendre leur utilisation plus fiable. DBpedia contient plus de 500 millions de triplets avec 60 000 rôles et plus de 480 000 concepts, sachant que nous avons ajouté un concept $\top$ subsumant tous les concepts de DBpedia sans concept parent, dont le concept `owl:Thing`. Il est intéressant de noter que l'analyse des résultats obtenus avec DBpedia montre que les contraintes découvertes contribuent non seulement à mieux caractériser les données mais aussi à mieux cerner les choix de conception et d'implantation de la base de connaissances.

C'est le cas également pour la deuxième base de connaissances sur laquelle nous avons fait tourner l'algorithme C3M, que nous appellerons COINS, qui est le résultat d'un processus d'intégration manuel mené dans le cadre du projet européen ARIADNE[(7)]. Ses créateurs ont utilisé l'ontologie CIDOC-CRM[(8)] pour intégrer les

[(7)] `http://ariadne-infrastructure.eu/`

[(8)] `http://www.cidoc-crm.org/`

contenus de 5 ressources numismatiques, construites par des institutions de différents pays européens [9]. COINS contient un peu plus de 3 millions de triplets avec 114 concepts et 373 rôles, relevant du CIDOC-CRM d'une part et d'ARIADNE d'autre part.

Nos algorithmes sont implémentés en Java, utilisent la bibliothèque de programmation pour RDF Jena[(9)] pour interroger la base de connaissances concernée. La base COINS, mise à notre disposition par ses concepteurs dans le triplestore Blazegraph (v2.1.4), est installée sur un serveur local ayant pour processeur un Dual Intel Xeon E5620 4 coeurs, dans une machine virtuelle sous Linux avec 32 GB de mémoire virtuelle. DBpedia est interrogée via son point d'accès SPARQL[(10)].

### 6.1. Résultats obtenus avec DBpedia

#### 6.1.1. *Impact du seuil minimum de cohérence* $\min_\tau$

Cette section s'intéresse au comportement de l'approche sur DBpedia notamment pour évaluer le passage à l'échelle de l'algorithme C3M. Pour cela, nous avons fixé le niveau de confiance $1 - \delta$ à 99 % et fait varier le seuil minimal de cohérence $\min_\tau$de $0,90$ à $0,99$ pour observer l'évolution de la collection des contraintes extraites.

La figure 6.1 reporte à gauche le temps d'exécution qui croît très rapidement lorsque le seuil de cohérence diminue. Cela s'explique par une augmentation très rapide de l'espace de recherche car les propriétés d'élagage s'avèrent moins efficaces. De ce fait, le nombre de contraintes extraites augmente également avec la diminution du seuil $\min_\tau$comme le montre le graphique de droite (figure 6.1) où est reporté l'augmentation du nombre de contraintes minimales de cardinalité maximale (nombre total est en traits continus, le nombre avec un contexte autre que $\top$ en grands pointillés et le nombre ayant une cardinalité maximale de 1 en pointillés fins). Il apparaît clairement que la majorité des contraintes ont une cardinalité maximale de 1. On constate également que la plupart dispose d'un contexte autre que $\top$, montrant bien l'utilité de notre approche.

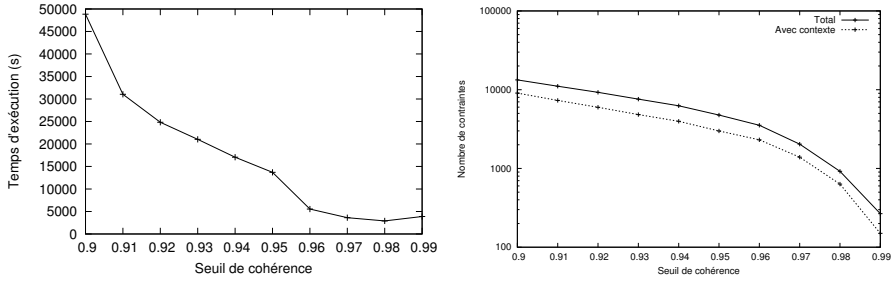


Figure 6.1. Evolution du temps d'exécution et du nombre de contraintes

La figure 6.2 donne la proportion de contraintes qui ont une cardinalité maximale égale à 1. Il est clair que la très grande majorité sont des contraintes de cardinalité

(9) `http://jena.apache.org`
(10) `https://dbpedia.org`

1. Cette proportion reste approximativement toujours la même. Cependant, la légère variation nous a été utile pour fixer un seuil minimal de cohérence égal à $0,97$ pour les expérimentations de la section suivante. En effet, nous avons observé empiriquement que les seuils inférieurs avaient tendance à identifier à tort des contraintes de cardinalité supérieure à 1. À l'inverse, au-delà de $0,97$, des contraintes de cardinalité maximale à 1 ont tendance à disparaître. Toujours sur la figure 6.2, la courbe de droite indique la proportion de contraintes non-minimales qui ont été supprimées (*i.e.*, nombre de contraintes supprimées divisé par le nombre de contraintes minimales ou non) en variant le seuil de cohérence. De manière intéressante, la réduction du nombre de contraintes grâce à la minimalité est importante quel que soit le seuil (entre 52 % et 65 %). Elle est moins forte aux alentours de $0,99$ mais beaucoup moins de contraintes sont alors identifiées. Pour rappel, ces contraintes élaguées ne sont pas informatives car redondantes avec d'autres plus générales. En d'autres termes, elles ne constituent aucun gain pour un système d'inférence et en plus, elles réduisent la lisibilité de l'extraction pour un utilisateur.

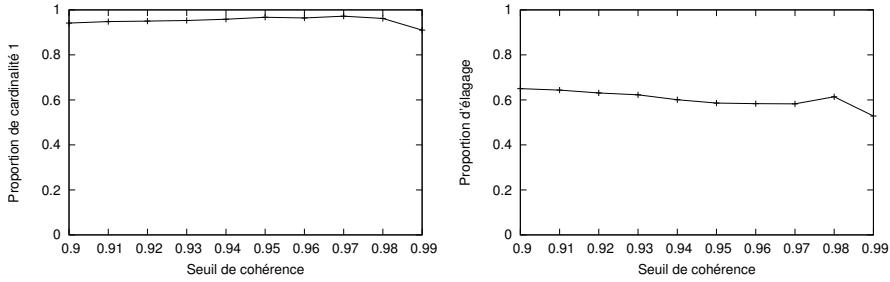


FIGURE 6.2. Evolution des proportions de contraintes de cardinalité 1 et élaguées

Pour évaluer la qualité des contraintes découvertes, nous avons annoté semi automatiquement un ensemble $C^*$ de 5041 contraintes, prises parmi les 13313 contraintes découvertes avec le seuil $\min_\tau = 0,90$. Par exemple, des rôles de la forme `birth`$X$ ont une cardinalité maximale de 1 pour tous les concepts. L'ensemble $C^*$ couvre 667 rôles et 2150 concepts. À partir de là, la précision d'un ensemble de contraintes $C$ correspond à la proportion de contraintes correctes par rapport au nombre de contraintes annotées ($C \cap C^*$). La figure 6.3 montre à gauche les précisions des ensembles de contraintes produits par C3M en fonction du seuil minimum de cohérence $\min_\tau$. La précision croît avec ce seuil et pour des seuils supérieurs ou égaux à $0,94$, elle est excellente puisqu'environ 95 % des contraintes sont correctes. Au-delà de $0,97$, il y a une décroissance marginale sans doute due au fait que des contraintes correctes ne sont pas retournées par C3M à des seuils aussi élevés. Notons que nous n'avons pas comparé notre approche avec la méthode proposée dans [20]. En effet, comme sur les exemples illustrés table 1.1, nous avons remarqué que la cardinalité maximale obtenue pour les 5 041 contraintes annotées était toujours incorrecte (d'une valeur supérieure à la valeur réelle attendue).

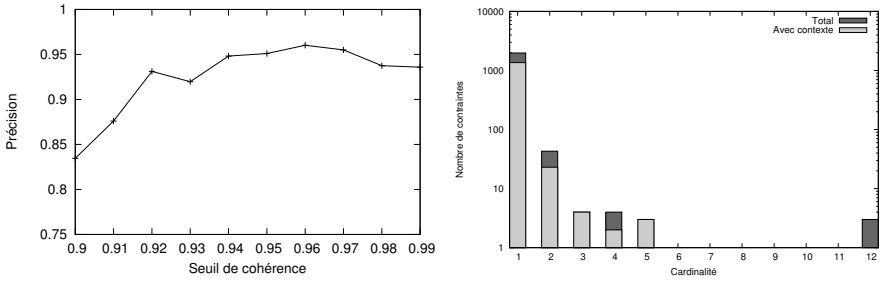


FIGURE 6.3. Précision des contraintes et répartition des contraintes suivant leur cardinalité avec et sans contexte

En résumé, notre approche passe bien à l'échelle sur DBpedia comportant environ 500 millions de triplets grâce aux méthodes d'élagage. La majorité des contraintes extraites ont un contexte, ce qui démontre l'intérêt d'explorer la hiérarchie des concepts de la base de connaissance. Enfin la précision des contraintes découvertes est très bonne, autour de 95 % pour $\min_\tau \geqslant 0{,}94$.

### 6.1.2. *Analyse qualitative*

Cette section s'intéresse à l'analyse d'une extraction des contraintes minimales de cardinalité maximale pour un seuil minimal de cohérence. Comme indiqué ci-avant, le seuil minimal de cohérence de 0, 97 a été retenu car il s'agit d'un bon compromis entre précision et rappel. La figure 6.3 présente à droite la répartition des 2 036 contraintes minimales de cardinalité maximale en fonction de la cardinalité. Les barres noires et grises représentent respectivement toutes les contraintes et celles avec un contexte différent de $\top$. Par exemple, il n'y a aucune contrainte de cardinalité maximale de 3 sans contexte.

Tout d'abord, cette répartition des contraintes montre que la très grande majorité (soit 1 979) sont des contraintes de cardinalité maximale égale à 1. Il peut paraître surprenant de disposer de 3 contraintes de cardinalité 12 alors qu'il n'y en a aucune avec une cardinalité comprise entre 6 et 11. En réalité, ces 3 contraintes résultent d'un biais de conception de DBpedia sur lequel nous revenons ci-après. De plus, environ la moitié des contraintes (soit 1 391) dispose d'un contexte soulignant l'importance de ne pas considérer uniquement le niveau $\top$. Par exemple, le rôle `country` décrit des entités très diverses au sein de DBpedia et aucune cardinalité ne peut être identifiée au niveau $\top$. En revanche, 81 classes ont une cardinalité maximale pour ce même rôle. Nous avons également noté que certaines contraintes apparaissent dans des contextes très précis, par exemple 3 contraintes existent pour des insectes apparaissant au niveau 10 de la hiérarchie des animaux.

Comme dit précédemment, l'ensemble de contraintes découvertes a une bonne précision. De façon intéressante, les contraintes incorrectes dénotent des biais de

TABLE 6.1. Exemples de contraintes minimales de cardinalité maximale

| Contexte $C$ | Rôle $R$ | Card. $M$ | Biais |
|---|---|---|---|
| $\top$ | `abstract` | 12 | conceptuel |
| `http://schema.org/School` | `country` | 2 | logique |
| $\top$ | `birthDate` | 2 | physique |

représentation, aux niveaux conceptuel, logique ou physique. Par exemple, au niveau conceptuel, la contrainte $\top \sqsubseteq (\leqslant 12$ `abstract`$)$ reflète un choix de DBpedia de renseigner majoritairement 12 langues. Ce choix de conception récurrent induit un biais sur la cardinalité observée. La méthode de représentation au niveau logique peut également influencer la cardinalité des rôles. Par exemple, la contrainte `http://schema.org/School` $\sqsubseteq (\leqslant 2$ `country`$)$ fixée à 2 peut surprendre car une école est localisée dans un unique pays. En observant les données, on constate que de nombreux établissements anglais sont rattachés à la fois à l'Angleterre et au Royaume-Uni. Il est clair qu'un seul rattachement à l'Angleterre nation constitutive du Royaume-Uni aurait été suffisant. Il y a eu un choix de « saturation » que l'on retrouve régulièrement avec les rôles transitifs. Enfin, un choix de modélisation physique peut même conduire à une modification de la cardinalité réelle. Par exemple, chaque personne a une date de naissance unique, nous identifions cependant une cardinalité de 2 car de nombreuses dates sont représentées avec deux formats distincts.

En résumé, pour un seuil de cohérence élevé $(0,97)$, notre méthode produit un ensemble raisonnable de 2036 contraintes contextuelles de cardinalité maximale. Les contraintes découvertes font sens car on retrouve des contraintes réelles ou induites par les choix de représentation.

### 6.2. EXPÉRIMENTATIONS AVEC LA BASE COINS

La base COINS concerne essentiellement des pièces de monnaies mais, par choix des intégrateurs, il n'y existe pas de concept *Coin*. Les individus correspondant à des pièces sont par exemple des instances de `E22_Man_Made_Object` caractérisées par certains URIs (ex. <`http://nomisma.org/id/coin`>) comme valeur objet de certains rôles (ex. `P2_has_type`). Plusieurs rôles et plusieurs URIs sont caractéristiques de pièces de monnaies, selon les différentes bases sources. Voulant vérifier que l'algorithme C3M produit des règles pertinentes pour la réalité représentée, en l'occurrence des pièces de monnaies, nous avons décidé de *construire la hiérarchie de concepts* fournie en entrée de C3M, de la façon suivante :

Au *premier niveau*, la hiérarchie construite contient tous les concepts $C_i$ de COINS, soit 114 concepts ($i \in [1..114]$). Tous ces concepts sont des sous-concepts du *concept racine* $\top$ *au niveau zéro*, *i.e.* pour tout $i$, nous avons $C_i \sqsubseteq \top$. Au *deuxième niveau* la hiérarchie contient tous les concepts $C_i^j$ définis par $C_i^j := C_i \sqcap (\exists R_j.\top)$ où $C_i$ ($i \in [1..114]$) et $R_j$ ($j \in [1..373]$) sont respectivement des concepts et rôles de COINS. À

ce niveau, 42 522 concepts $C_i^j$ sont ainsi définis. Enfin, au *troisième niveau*, la hiérarchie contient tous les concepts $C_i^{j,k}$ définis par $C_i^{j,k} := C_i \sqcap (\exists R_j.\{a_k\})$ où $C_i$ ($i \in [1..114]$) et $R_j$ ($j \in [1..373]$) sont respectivement des concepts et rôles de COINS, et $a_k$ est un individu du co-domaine de $R_j$, *i.e.* $a_k \in (\exists R_j^{-1}.\top)$. Grâce à ce dernier niveau, il est possible de considérer des contextes correspondant à des ensembles de pièces de monnaies[(11)]. Notons finalement que pour tout $i, j, k$, nous avons $C_i^{j,k} \sqsubseteq C_i^j \sqsubseteq C_i$. Globalement, cette hiérarchie comporte 3 160 357 contextes, donc pour les 373 rôles de COINS cela représente plus d'un milliard de contraintes contextuelles possibles (exactement 1 178 813 161 contraintes). Néanmoins, l'utilisation des propriétés 5.1 et 5.2 permet d'élaguer une grande partie de cet espace de recherche.

Tous les résultats présentés dans cette section ont été obtenus avec un *seuil minimal de confiance* $1 - \delta = 99$ % (pour des contraintes les plus certaines possibles) et un *seuil minimal de cohérence* $\min_\tau = 0,95$ (pour des contraintes suffisamment probables). Ce seuil a été défini expérimentalement. Comme indiqué dans la section suivante, sur des bases de connaissances de plus grande taille comme DBpedia, un seuil plus élevé est préférable. Pour autant l'approche est relativement peu sensible aux seuils.

### 6.2.1. *Analyse quantitative*

Avec ces paramètres, la propriété 5.1 nous indique qu'une contrainte $C \sqsubseteq (\leqslant M\ R)$ ne peut être suffisamment certaine si son contexte $C$ contient moins de $\alpha = \frac{\log(1/\delta)}{2(1-\min_\tau)^2} = 922$ instances. Ainsi, l'utilisation de la propriété 5.1 permet de n'explorer que 16 641 contraintes, soit moins de 0,002 % des plus de 1 milliard de contraintes possibles. Qui plus est, notre expérience montre que la propriété 5.2 permet de réduire encore de 82,5 % la taille de l'espace de recherche à explorer. Au final, avec les seuils choisis notre algorithme ne cherche à détecter une cardinalité maximale que pour 2 909 contextes, avec un temps total de calcul de moins de 50 minutes.

La table 6.2 donne une vue globale et quantitative du résultat de l'exploration réalisée. Sur les 2 909 contraintes contextuelles possibles, notre algorithme a détecté au total 887 contraintes de cardinalité maximale, 595 d'entre elles étant des contraintes minimales. Sur cet exemple, le critère de minimalité permet donc de réduire de près de 67 % le nombre de contraintes retournées. On constate que les contraintes les plus nombreuses sont des cardinalités maximales avec $M = 1$, ce qui correspond à des contraintes où pour un rôle donné $R$, tout sujet $s$ est en relation avec au plus un objet $o$. Néanmoins de très nombreuses contraintes sont trouvées avec des cardinalités maximales $M \in \{2, 3\}$ (37 % des contraintes minimales détectées). On note également que si des contraintes de cardinalités maximales sont détectées dès le niveau 0 (65 contraintes avec un contexte $C \equiv \top$), la recherche de contraintes contextuelles est particulièrement pertinente. Il faut en effet noter que les contraintes les plus nombreuses sont trouvées au niveau 3 (75 % des contraintes détectées), sachant que par construction, c'est à ce niveau de la hiérarchie construite que sont caractérisées les pièces de monnaie.

(11) `E22_Man_Made_Object` $\sqcap \exists$ `P2_has_type`. {<http://nomisma.org/id/coin>} par exemple.

TABLE 6.2. Répartition par niveau et cardinalité maximale $M$ des contraintes minimales détectées

| | **Niveau dans la hiérarchie** | | | | |
|---|---|---|---|---|---|
| | **0** | **1** | **2** | **3** | **Total** |
| $M$ | $\top$ | $C_i$ | $C_i^j$ | $C_i^{j,k}$ | |
| 1 | 60 | 28 | 10 | 222 | 320 |
| 2 | 3 | 6 | 9 | 90 | 108 |
| 3 | 0 | 7 | 14 | 92 | 113 |
| 4 | 1 | 0 | 8 | 20 | 29 |
| 5 | 1 | 0 | 0 | 16 | 17 |
| 6 | 0 | 0 | 0 | 8 | 8 |
| **Total** | **65** | **41** | **41** | **448** | **595** |

6.2.2. *Analyse qualitative*

Tout d'abord, dès le niveau 0, notre méthode permet de retrouver des contraintes fonctionnelles attendues, par exemple pour les 3 rôles du CIDOC-CRM : `P1_is_identified_by`, `P52_has_current_owner`, `P50_has_current_keeper`. De telles contraintes indiquent que si un sujet décrit dans COINS possède plus d'un identifiant, un propriétaire ou un conservateur, alors on peut en déduire que ces identifiants (respectivement, propriétaires et conservateurs) sont identiques. Concernant le rôle `P45_consists_of` du CIDOC-CRM (permettant de décrire les matériaux constitutifs d'un objet), il est intéressant de noter qu'une cardinalité maximale de 2 est détectée dès le niveau 1 pour la classe `E22_Man_Made_Object`. La base de connaissances décrit notamment des médailles constituées d'or et de pierre précieuse (telle l'agate). Pour ce même rôle, une cardinalité maximale de 1 est détectée au niveau 3 pour les pièces de monnaie. Un même type de contrainte (avec $M = 1$) est trouvée au niveau 3 pour tous les contextes décrivant des pièces, concernant le rôle `P62_depicts` (ce qui est dépeint sur l'objet). C'est raisonnable car dans le cas d'une pièce de monnaie, on trouve le plus souvent une seule représentation figurative (sur une des deux faces de la pièce), alors qu'une telle contrainte n'est pas valide pour d'autres objets.

En résumé, l'étude de l'ensemble des contraintes extraites par notre méthode a mis en évidence des redondances dans la base, du fait de choix d'intégration. Dans une phase de post-traitement, la connaissance de telles redondances pourrait réduire encore le nombre de contraintes extraites.

## 7. CONCLUSION

Nous avons présenté la première proposition de calcul de contraintes de cardinalité maximale dans une base de connaissances du web sémantique qui produit des contraintes significatives par rapport à la réalité. Ces grandes bases de connaissances, reflet d'une intelligence collective, sont générées à partir de l'expertise limitée de

nombreux contributeurs et souffrent encore, tantôt de lacunes dans les informations, tantôt d'incohérences. Utiliser leurs contenus courants afin de mieux caractériser les connaissances représentées est donc très utile, comme montré dans l'état de l'art : cela permet aux applications qui exploitent ces grandes bases de connaissances de produire des résultats plus fiables.

Nos expérimentations démontrent la faisabilité d'une exploration systématique de grandes bases de connaissances telle que DBpedia (plus de 500 millions de triplets) pour la recherche de contraintes contextuelles de cardinalité maximale grâce à l'algorithme C3M que nous proposons dans cet article. Les propriétés utilisées par C3M réduisent drastiquement le nombre de contraintes qui sont obtenues, ce qui rend possible leur examen manuel. Cela nous a permis de vérifier que ces contraintes sont pertinentes une fois identifiés les bons contextes. Les expérimentations sur la base COINS montrent la généricité de C3M, qui peut prendre comme hiérarchie de contextes aussi bien la hiérarchie définie dans la base de connaissances $\mathcal{K}$ analysée, qu'une hiérarchie construite à partir de $\mathcal{K}$. Comme c'est le cas également avec DBpedia, ces expérimentations démontrent l'importance du contexte dans cette découverte de contraintes.

Nous avons pour perspective d'exploiter les contraintes extraites pour calculer la confiance associée à des règles découvertes dans la base de connaissances considérée. Mais avant cela, nous souhaiterions étendre notre approche aux contraintes contextuelles de cardinalité minimale et bénéficier davantage des capacités de raisonnement. Pour l'instant, nous tenons compte uniquement de la hiérarchie des classes pour réduire l'ensemble de contraintes avec le rôle `rdfs:subClassOf`, donc C3M s'applique pour tout formalisme utilisé pour décrire la base de connaissance (RDFS, ou n'importe quel OWL). Nous pourrions améliorer l'approche en exploitant par exemple les rôles `owl:sameAs` (comme dans [20]), `owl:equivalentClass` et `owl:equivalentProperty`, s'ils sont présents. Enfin, les contraintes de cardinalités découvertes pourraient être des contraintes de cardinalité qualifiées (OWL 2 EL), munies d'un taux de cohérence.

**Remerciements**

Les auteurs remercient Corentin Viemon pour sa contribution à l'implémentation en Java des propositions contenues dans cet article, au cours de son projet de Master 2 BDMA à Blois.

**Bibliographie**

[1] E. A. S. Aly, M. L. Diakité, A. Giacometti, B. Markhoff & A. Soulet, « Découverte de cardinalité maximale contextuelle dans les bases de connaissances », in *Actes de la Conférence Nationale d'Intelligence Artificielle et Rencontres des Jeunes Chercheurs en Intelligence Artificielle (CNIA+RJCIA 2018), Nancy, France*, 2018, p. 86-93.

[2] M. Atencia, J. David & F. Scharffe, « Keys and Pseudo-Keys Detection for Web Datasets Cleansing and Interlinking. », in *Proc. of the 18th International Conference on Knowledge Engineering and Knowledge Management*, Springer, 2012, p. 144-153.

[3] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak & Z. Ives, « Dbpedia : A Nucleus for a Web of Open Data », in *The semantic web*, Springer, 2007, p. 722-735.

[4] F. Baader, D. Calvanese, D. L. McGuinness, D. Nardi & P. F. Patel-Schneider (éds.), *The Description Logic Handbook : Theory, Implementation, and Applications*, Cambridge University Press, New York, NY, USA, 2003.

[5] F. Baader & U. Sattler, « Expressive number restrictions in description logics », *Journal of logic and computation* **9** (1999), nº 3, p. 319-350.

[6] F. Darari, W. Nutt, G. Pirrò & S. Razniewski, « Completeness Statements about RDF Data Sources and Their Use for Query Answering », in *Proc. of International Semantic Web Conference* (Berlin, Heidelberg), Springer, 2013, p. 66-83.

[7] F. Darari, S. Razniewski, R. E. Prasojo & W. Nutt, « Enabling Fine-Grained RDF Data Completeness Assessment », in *Proc. of International Conference on Web Engineering* (Cham), Springer International Publishing, 2016, p. 170-187.

[8] F. Erxleben, M. Günther, M. Krötzsch, J. Mendez & D. Vrandečić, « Introducing Wikidata to the linked data web », in *Proc. of International Semantic Web Conference*, Springer, 2014, p. 50-65.

[9] A. Felicetti, P. Gerth, C. Meghini & M. Theodoridou, « Integrating Heterogeneous Coin Datasets in the Context of Archaeological Research », in *Proc. of the Workshop on Extending, Mapping and Focusing the CRM, co-located with 19th ICTPDL conference*, CEUR-WS.org, 2015, p. 13-27.

[10] L. A. Galárraga, K. Hose & S. Razniewski, « Enabling Completeness-aware Querying in SPARQL », in *Proc. of the 21st Workshop on the Web and Databases*, ACM, 2017, p. 19-22.

[11] L. A. Galárraga, S. Razniewski, A. Amarilli & F. M. Suchanek, « Predicting completeness in knowledge bases », in *Proc. of the 10th ACM International Conference on Web Search and Data Mining*, ACM, 2017, p. 375-383.

[12] L. A. Galárraga, C. Teflioudi, K. Hose & F. Suchanek, « AMIE : Association Rule Mining Under Incomplete Evidence in Ontological Knowledge Bases », in *Proc. of World Wide Web Conference*, ACM, 2013, p. 413-422.

[13] W. Hoeffding, « Probability inequalities for sums of bounded random variables », *Journal of the American Statistical Association* **58** (1963), nº 301, p. 13-30.

[14] J. Lajus & F. M. Suchanek, « Are all people married ? Determining obligatory attributes in knowledge bases », in *Proc. of World Wide Web conference*, 2018, p. 1115-1124.

[15] S. W. Liddle, D. W. Embley & S. N. Woodfield, « Cardinality constraints in semantic data models », *Data & Knowledge Engineering* **11** (1993), nº 3, p. 235-270.

[16] H. Mannila & H. Toivonen, « Levelwise search and borders of theories in knowledge discovery », *Data Mining and Knowledge Discovery* **1** (1997), nº 3, p. 241-258.

[17] P. Mirza, S. Razniewski, F. Darari & G. Weikum, « Enriching Knowledge Bases with Counting Quantifiers », in *Proc. of International Semantic Web Conference*, Springer, 2018, p. 179-197.

[18] A. Motro, « Integrity = Validity + Completeness », *ACM Transactional Database Systems* **14** (1989), nº 4, p. 480-502.

[19] E. Muñoz, « On learnability of constraints from RDF data », in *Proc. of International Semantic Web Conference*, Springer, 2016, p. 834-844.

[20] E. Muñoz & M. Nickles, « Mining cardinalities from knowledge bases », in *Proc. of International Conference on Database and Expert Systems Applications*, Springer, 2017, p. 447-462.

[21] N. Pernelle, F. Saïs & D. Symeonidou, « An automatic key discovery approach for data linking », *Web Semantics : Science, Services and Agents on the World Wide Web* **23** (2013), p. 16-30.

[22] S. Razniewski, F. Korn, W. Nutt & D. Srivastava, « Identifying the Extent of Completeness of Query Answers over Partially Complete Databases », in *Proc. of the ACM SIGMOD International Conference on Management of Data*, ACM, 2015, p. 561-576.

[23] S. Razniewski, F. Suchanek & W. Nutt, « But What Do We Actually Know ? », in *Proc. of the 5th Workshop on Automated Knowledge Base Construction*, 2016, p. 40-44.

[24] S. Shalev-Shwartz & S. Ben-David, *Understanding Machine Learning : From Theory to Algorithms*, Cambridge University Press, 2014.

[25] A. Soulet, A. Giacometti, B. Markhoff & F. M. Suchanek, « Representativeness of Knowledge Bases with the Generalized Benford's Law », in *Proc. of International Semantic Web Conference*, Springer, 2018, p. 374-390.

[26] C. Soutou, « Relational database reverse engineering : algorithms to extract cardinality constraints », *Data & Knowledge Engineering* **28** (1998), nº 2, p. 161-207.

[27] D. Symeonidou, V. Armant, N. Pernelle & F. Saïs, « SAKey : Scalable almost key discovery in RDF data », in *In Proc. of International Semantic Web Conference*, Springer, 2014, p. 33-49.

[28] D. Symeonidou, L. A. Galárraga, N. Pernelle, F. Saïs & F. Suchanek, « VICKEY : Mining Conditional Keys on Knowledge Bases », in *Proc. of International Semantic Web Conference*, Springer, 2017, p. 661-677.

[29] T. P. Tanon, D. Stepanova, S. Razniewski, P. Mirza & G. Weikum, « Completeness-Aware Rule Learning from Knowledge Graphs », in *Proc. of International Semantic Web Conference*, Springer, 2017, p. 507-525.

[30] B. Thalheim, « Fundamentals of cardinality constraints », in *Proc. of International Conference on Conceptual Modeling*, Springer, 1992, p. 7-23.

[31] J. Völker & M. Niepert, « Statistical schema induction », in *Proc. of Extended Semantic Web Conference*, Springer, 2011, p. 124-138.

[32] G. Weikum, J. Hoffart & F. M. Suchanek, « Ten Years of Knowledge Harvesting : Lessons and Challenges », *IEEE Data Engineering Bulletin* **39** (2016), nº 3, p. 41-50.

[33] D. Yeh, Y. Li & W. Chu, « Extracting entity-relationship diagram from a table-based legacy database », *Journal of Systems and Software* **81** (2008), nº 5, p. 764-771.

[34] A. Zaveri, A. Rula, A. Maurino, R. Pietrobon, J. Lehmann & S. Auer, « Quality assessment for linked data : A survey », *Semantic Web journal* **7** (2016), nº 1, p. 63-93.

---

Abstract. — Big semantic web knowledge bases (KB) are generated from collaborative platforms or by integration of various sources. This naturally induces lack of information, and inconsistencies. Moreover, missing data must not be considered as non existing data. Applications that query these KB's content need complementary information to decide whether the queried data is complete. Based on KB's volume, it is possible to discover such kind of information. We present an algorithm for extracting significant maximum cardinality rules from a knowledge base. We use Hoeffding's inequality to define the likelihood for a constraint to be significant. Experiments conducted on DBpedia and on a numismatic knowledge base resulting from an integration process demonstrate the feasibility of the approach and the relevance of the discovered contextual constraints.

Keywords. — Cardinality Mining, Contextual Constraints, Knowledge Base.

---

*Manuscrit reçu le 1er mars 2018, révisé le 8 février 2019, accepté le 8 février 2019.*