Où sont les termes ?

Béatrice Markhoff, Arnaud Soulet

Où sont les termes ?

Béatrice Markhoff[1], Arnaud Soulet[1]

[1] Université de Tours, LIFAT

prenom.nom@univ-tours.fr

Résumé

Pour donner une idée concise du contenu d'un graphe de connaissances, il est classique de montrer les classes et les propriétés qui y sont instanciées. Pourtant d'autres éléments peuvent informer sur ce contenu autant que les classes et les propriétés, ce sont les termes de vocabulaires contrôlés, des mots associés à des concepts. Nous présentons une étude sur le rôle et la place de ces termes dans des graphes de connaissances du Web, en particulier ceux conçus avec le CIDOC CRM. Nous expliquons les difficultés qu'il y a à les retrouver automatiquement. Nous recensons des requêtes simples pour ce faire, nous en proposons une plus complexe et nous présentons des résultats d'expérimentations sur des points d'accès SPARQL dans le domaine du patrimoine culturel, lesquels montrent que... la question reste ouverte.

Mots-clés

CIDOC CRM, graphe de connaissances, ontologie, SPARQL, thésaurus, terme, terminologie.

Abstract

To give a quick idea of a knowledge graph content, it is usual to show the classes and properties it instantiates. However, controlled vocabulary's terms can also inform about its content, as much as the classes and properties. We present a study on the role and the place of these terms in knowledge graphs, in particular those designed with the CIDOC CRM. We explain the difficulties of finding them automatically. We identify simple queries to do it, we propose a more complex one, and we present experimental results on SPARQL access points, in the cultural heritage domain. Those results show that... the question remains open.

Keywords

CIDOC CRM, knowledge graph, ontology, SPARQL, thesaurus, term.

1 Introduction

L'origine de cet article est un travail [3] sur le profilage de graphes de connaissances du Web se rapprochant de ce qui est proposé dans [16] ou [6]. Un profil consiste grossomodo en un graphe des classes et des propriétés instanciées. Par exemple pour le graphe de connaissances d'Epicherchel, contenant les objets sur lesquels ont été trouvées des inscriptions antiques à Césarée de Maurétanie, la figure 1 [1] montre dans sa partie gauche qu'il contient des instances de la classe *S19 Encounter event* et de la classe *E22 Man-Made Object* (en passant la souris sur le nom des classes on peut savoir combien) et également qu'il y a 182 triplets (sujet, prédicat, objet) où le sujet est de la classe *S19 Encounter event*, l'objet de la classe *E22 Man-Made Object* et le prédicat est la propriété *O19 has found object* [2]. Générer le profil d'un graphe de connaissances est utile pour informer sur son contenu et pour guider son exploration et son interrogation.

Pour Epicherchel comme pour les autres graphes de connaissances du portail OpenArchaeo [3], certaines propriétés ont pour objet un concept d'un vocabulaire contrôlé, dont le terme associé informe plus précisément sur le contenu du graphe que les seules classes et propriétés. Pour Epicherchel on voit en figure 1 que c'est le cas des propriétés *P2 has type*, *P101 had as general use* et *P45 consists of*. Par exemple *P101 had as general use* relie des instances de *E22 Man-Made Object* à des instances de *E55 Type* et aussi à des concepts, représentés par le noeud *autel et al*, concepts dont les termes associés sont listés lorsqu'on passe la souris sur ce noeud, comme montré dans la partie droite de la figure. Il est très utile de montrer ces termes dans un profil parce que savoir qu'il y a des milliers d'instances de *E22 Man-Made Object* dans un graphe indique juste que ce graphe contient des informations sur des objets du patrimoine culturel, mais voir en plus ces termes donne une idée plus précise, permettant de savoir dans l'exemple que ces objets relèvent de fouilles archéologiques de sites antiques du pourtour méditerranéen.

La problématique que nous montrons dans cet article est qu'il est difficile de caractériser ce genre de termes de vocabulaires contrôlés dans un graphe de connaissances et donc de les détecter automatiquement.

Pourtant ces termes sont particulièrement porteurs de connaissances : encore aujourd'hui et depuis l'origine de leur discipline, les archéologues apportent une attention particulière aux termes utilisés pour renseigner les bases de données recensant leurs découvertes. Des typologies se sont constituées, motivées par la nécessité de nommer «

1. Visible en ligne ici : `https://kgsumviz.univ-tours.fr/Home.php`
2. Ces classes et propriétés appartiennent au CIDOC CRM ou à ses extensions : `https://cidoc-crm.org`
3. `http://openarchaeo.huma-num.fr/explorateur/home`

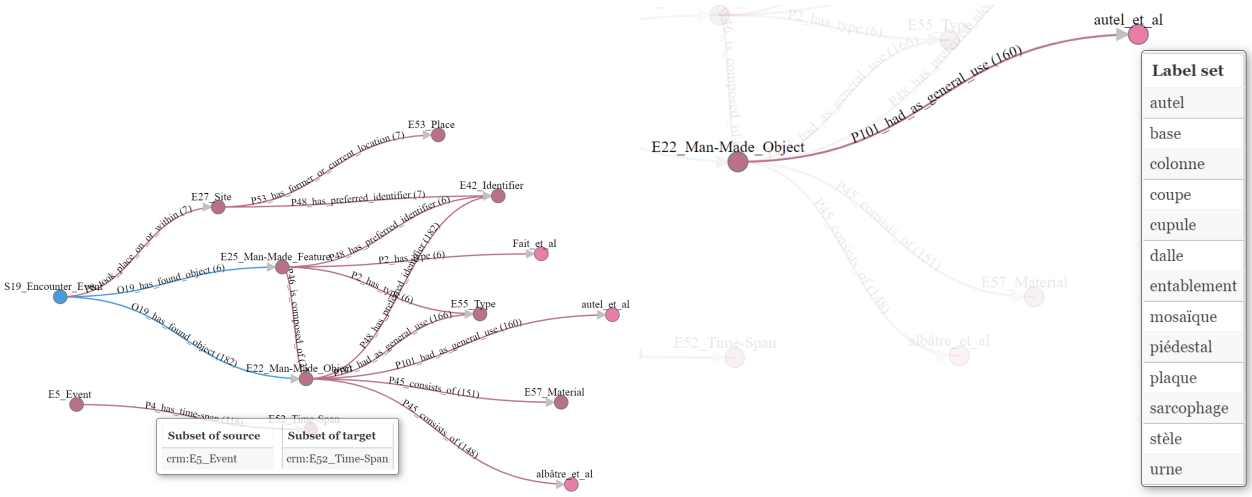

FIGURE 1 – Profil [3] du graphe Epicherchell dans OpenArchaeo.

correctement » les objets découverts, leurs fonctions, leur forme, leur constitution (matériaux), et toutes leurs caractéristiques. Il existe une tension récurrente entre la liberté de choix et de précision des termes utilisés, choix et précision qui relèvent de l'activité savante, et la nécessité de consensus, au moins à un assez haut niveau d'abstraction, pour permettre le partage d'information [10]. Le besoin d'une normalisation des termes utilisés dans les descriptions est d'autant plus ressenti avec l'ouverture des ressources via le Web, pour ce qu'il permet d'échanges et de partage. Comme l'interopérabilité sémantique passe en particulier par les termes utilisés, la communauté a mis en œuvre des systèmes en ce sens [4, 13], en parallèle de travaux sur la représentation des connaissances sous la forme d'ontologies. De façon intéressante, en archéologie comme ailleurs (par exemple dans le domaine médical avec le système d'indexation MeSH et le vocabulaire SNOMED CT), le travail sur les termes est pris en charge par des spécialistes de la gestion de l'information (documentalistes) et par des spécialistes du domaine, en parallèle ou conjointement, avec cette motivation double : d'une part *indexer pour retrouver* l'information, d'autre part *décrire* son objet d'étude et son étude elle-même. Nous revenons sur ces deux motivations distinctes dans la section 2, en donnant également un aperçu de l'importance de la production de ressources terminologiques dans le Web sémantique.

Ainsi, alors que les ontologies du Web sémantique décrivent déjà les domaines représentés via un vocabulaire structuré (Terminological Box des logiques de description), les termes sont très souvent *aussi dans les données* (Assertional Box) et c'est même une bonne pratique recommandée dans [12] : il faut utiliser des termes de vocabulaires partagés, de préférence standardisés, pour encoder les données et les métadonnées. Les auteurs de la recommandation indiquent que les bénéfices de cette bonne pratique sont : la réutilisation, la facilité de traitement, la compréhension, la confiance et l'interopérabilité. Dans ces vocabulaires les termes sont associés à des URIs, décrits de façon non ambiguë et qui peuvent avoir des labels dans différentes langues. Nous verrons en section 2 que ces vocabulaires peuvent prendre la forme d'ontologies (décrites en RDFS ou OWL) ou de thésauri (décrits en SKOS pour la plupart). Etant donné que les ontologies sont également, comme les ressources terminologiques, des supports d'interopérabilité sémantique, nous rappelons en section 3 les relations qu'elles peuvent entretenir avec les ressources terminologiques, avec un focus sur l'ontologie CIDOC CRM [2] pour le patrimoine culturel.

Ces réflexions ont pour objet de déterminer *comment détecter des termes* dans un graphe de connaissances[4]. Car autant la propriété `rdf:type` dénote une instance de classe et la 2ème position dans un triplet dénote une instance de propriété, autant détecter un terme ne repose sur aucun élément de syntaxe aussi simple. Dans la section 4, nous listons donc différentes formes de syntaxe pouvant dénoter des termes et les requêtes SPARQL correspondantes, puis nous présentons des résultats d'expérimentations sur différents points d'accès SPARQL liés au patrimoine culturel, montrant dans quelle mesure ces indices fonctionnent. Enfin, nous concluons en section 5 sur cette étude.

2 Que sont les termes dans le Web sémantique

Ce que nous appelons ressource terminologique dans le Web sémantique est conçu selon deux objectifs, l'indexation à des fins de recherche d'information et l'élaboration de terminologies.

2.1 Terminologies

Lorsqu'une communauté recherche un consensus sur les mots à utiliser pour décrire des éléments de son domaine, y compris dans des langues différentes, elle conçoit une terminologie. C'est utile en particulier pour renseigner de fa-

4. Graphe de connaissances au sens défini dans [7] qui, de façon intéressante, ne parle pas du tout de termes.

çon cohérente des champs de bases de données. En ce sens les termes sont des *données*. La norme ISO 1087 :2019 du groupe ISO/TC 37 définit une terminologie comme un ensemble de désignations utilisées dans un langage de spécialité, où une désignation représente un *concept* par un *signe* qui le dénote. L'ISO/TC 37 est également à l'origine des standards TMF (Terminological Mark-up Framework) et LMF (Lexical Mark-up Framework) qui ont inspiré l'ontologie OntoLex-lemon [5] (représentation des propriétés morpho-syntaxiques des entrées lexicales et de leur sens) et son extension pour la terminologie en cours de définition, Termlex [6], dédiée à la documentation des informations sur les termes. Cette proposition permet de représenter clairement l'interface entre syntaxe (signe) et sémantique à l'aide du concept LexicalSense, qui peut faire le lien vers une ontologie dans laquelle le concept est décrit [7]. La question de ce lien entre forme lexicale et concept pour la description de termes fait également l'objet de proposition de système onto-terminologique [14].

Cependant en général dans le Web, certaines terminologies sont réalisées juste sous forme d'ontologie, comme le Dublin Core [8], et d'autres sont réalisées juste avec SKOS ou SKOS-XL [9], comme le AAT du Getty [10]. Pourtant SKOS a une expressivité limitée pour cet usage, puisque c'est une ontologie pour définir des thésauri, taxonomies, schémas de classification ou systèmes de vedettes-matières, utilisés dans des systèmes documentaires à des fins d'indexation et de recherche d'information.

2.2 Indexation et recherche d'information : les thésauri

Un thésaurus est un vocabulaire contrôlé et structuré dans lequel les concepts sont représentés par des termes, où des relations entre les concepts sont explicitées et où les termes préférés sont accompagnés d'entrées de synonymes ou de quasi-synonymes (voir ISO 25964-1 sections 2.62 thésaurus et 2.35 thésaurus multilingue). Dans ce cadre un concept est une unité de pensée et un terme est un mot ou une expression utilisée pour étiqueter un concept (voir ISO 25964-1 sections 2.11 Concept et 2.61 Terme). Ces définitions se rapprochent de celles d'une terminologie et justifient le terme de « ressource terminologique » pour parler aussi bien de terminologie que de thésaurus. Une différence entre thésauri et terminologies réside toutefois dans leur vocation (nous avons vu que celle des terminologies est le consensus sur les désignations utilisées dans un langage de spécialité, soit l'interopérabilité sémantique). L'objectif principal des thésauri est d'indexer et de retrouver des éléments (souvent des documents) en fonction de leur contenu : « le document D traite le sujet C ». Le thésaurus sert alors de structure d'accès, sachant que les déclarations de synonymie entre termes d'une part, et d'autre part la relation hiérarchique entre concepts, permettent au système de recherche d'information d'élargir ou de restreindre les requêtes. Dans ce but, la hiérarchie utilisée dans un thésaurus couvre la relation de subsomption, la relation de partition, parfois aussi la relation d'instanciation, fusionnées en une relation hiérarchique unique dans certains thésaurus pour répondre à cette définition fondée sur l'usage : « le concept A est plus large que le concept B » si « dans toute recherche de A, les articles traitant de B devraient être retournés ». Les grands thesaurus sont organisés en facettes, qui regroupent des hiérarchies de concepts pour faciliter la recherche d'information. Comme le note [9], la structure des thésauri n'est pas utile à des raisonnements plus généraux, contrairement aux ontologies.

2.3 Ressources terminologiques dans le Web

On trouve des ressources terminologiques sous la forme de thésauri ou d'ontologies dans le Web sémantique, les deux étant supports d'interopérabilité sémantique, les premiers au niveau données et les secondes au niveau métadonnées. Les ressources terminologiques sont plus massivement réutilisées que des ontologies conçues pour représenter un domaine de connaissance, dans la mesure où le consensus nécessaire à leur réutilisation porte sur les termes (et leur définition en contexte), et non pas sur la question plus complexe de la manière dont la représentation du domaine est organisée et structurée (ontologie). Il est plus simple de choisir un terme pertinent dans un thésaurus que de comprendre et réutiliser une ontologie, sauf pour les plus simples d'entre elles comme FOAF et le Dublin Core, qui sont des ressources terminologiques. Ainsi, il existe de nombreuses et, pour certaines, très grandes ressources terminologiques dans le Web. Par exemple dans le domaine biomédical UMLS [11] rassemble des concepts de plusieurs dizaines de terminologies, MeSH [12] (défini par la bibliothèque nationale de médecine des Etats-Unis) permet d'indexer les répertoires d'articles Medline et PubMed, quand SNOMED CT [13] rassemble plusieurs centaines de milliers de concepts utilisés dans des environnements cliniques, en particulier pour les dossiers patients. De même dans le domaine environnemental, AGROVOC [14] regroupe plus de 38 000 concepts de l'alimentation, agriculture, pêche, foresterie, etc. auxquels sont associés plus de 800 000 termes dans 40 langages. Dans le domaine du patrimoine culturel, le Backbone thesaurus de DARIAH [15] est une initiative pour l'agrégation et la maintenance de vocabulaires construits dans des communautés, comme les PACTOLS déjà cités pour l'archéologie, mais c'est surtout le vocabulaire AAT

5. https://www.w3.org/2016/05/ontolex/
6. https://www.w3.org/community/ontolex/wiki/Terminology
7. Voir http://anr-sesames.map.cnrs.fr/onto/shart/index.html pour un exemple d'utilisation
8. Voir https://www.dublincore.org/specifications/dublin-core/dcmi-terms/
9. https://www.w3.org/TR/skos-reference/
10. Art and Architecture Thesaurus : https://www.getty.edu/research/tools/vocabularies/aat/
11. Unified Medical Language System : https://www.nlm.nih.gov/research/umls/index.html
12. Medical Subject Headings : https://www.ncbi.nlm.nih.gov/mesh
13. Systematized Nomenclature Of Medicine Clinical Terms : https://www.snomed.org/
14. https://www.fao.org/agrovoc/
15. https://www.backbonethesaurus.eu/

du Getty, déjà cité également, qui est utilisé et avec lequel les thésauri locaux sont alignés, par exemple dans EUROPEANA [16] ou dans la plateforme ARIADNE+ [17]. Celle-ci organise les recherches possibles selon trois axes Quand-Où-Quoi : le Getty AAT est utilisé dans l'axe Quoi, pour décrire ce qui est recherché [18]. Il est intéressant de noter ici que pour des axes de recherche comme Quand (périodes historiques), Où (lieux) et Qui (personnes, organisations), aussi bien dans la plateforme ARIADNE+ que dans OpenArchaeo, des URIs issus de listes d'autorité sont également utilisés, mais il s'agit alors *d'entités nommées*, qui se réfèrent chacune à un élément unique, et non plus des *termes* qui, eux, *sont des « universaux » [11] au même titre que les classes et les propriétés d'une ontologie*, même s'ils sont des instances de `skos:Concept`. Cette dernière remarque sera exploitée dans la section 4.

3 Relations avec les ontologies

3.1 Ontologie versus thésaurus

Il n'est pas question ici de définir précisément ce qu'est une ontologie dans le Web comme le font les auteurs de [9], rappelons simplement que c'est un modèle formel, et consensuel, d'une conceptualisation d'un domaine de connaissances. Il est constitué de description intensionnelle (TBox) et extensionnelle (ABox, rassemblant les instances). Il comporte un ensemble d'entités, de relations et d'axiomes pour les décrire. Une entité y est vue comme une classe qui a des instances, le lien de subsomption étant que si A subsume B alors toute instance de B est aussi instance de A. Les relations sont décrites par leur domaine et codomaine et il y a également un lien de subsomption entre relations. Selon les besoins d'autres précisions peuvent être spécifiées. L'ensemble de ces déclarations est exploitable automatiquement par des raisonneurs. Ainsi, une bonne pratique recommandée dans [5] pour modéliser des ressources et leurs contenus est d'affecter à la ressource une propriété `hasTopic` dont l'objet est instance d'une classe représentant ce contenu. Sachant que les classes représentant les contenus sont organisées selon leurs liens de subsomption, cela permet de bénéficier directement du raisonneur : demander des ressources portant sur un métal revient à demander aussi celles portant sur toutes les sortes de métal subsumées par la classe Metal. On reconnaît là la vocation d'un thésaurus, mais dans un thésaurus c'est un système ad hoc qui réalise cette généralisation de la requête, pas un raisonneur générique. La question de transformer les connaissances contenues dans un thésaurus en ontologie pour les exploiter automatiquement par un raisonneur est abordée de longue date [1, 8]. Elle n'est pas simple, notamment du fait que la relation broader-narrower d'un thésaurus peut être aussi bien une relation de subsomption qu'une relation de partition, et que la relation associative générique souvent présente dans un thésaurus est également complexe à transposer automatiquement dans une ontologie. De plus, il faut décider quels concepts du thésaurus sont représentés par des classes et lesquels sont des données (instances). Pour autant il existe des versions de SNOMED CT et d'AGROVOC sous forme d'ontologies OWL. Dans le domaine du patrimoine culturel, la communauté qui définit et maintient l'ontologie CIDOC CRM a une politique claire pour l'usage de descriptions terminologiques conjointement avec l'ontologie, que nous présentons succinctement dans la section suivante.

3.2 CIDOC CRM

CIDOC CRM est une ontologie conçue pour supporter l'interopérabilité sémantique de ressources numériques du patrimoine culturel, administrée par un groupement d'intérêts (SIG) depuis les années 80. La question des termes est abordée dès l'introduction du document qui la définit, avec ce titre : « About Types ». Une classe particulière, `E55 Type`, est destinée à regrouper les termes de thésaurus et vocabulaires contrôlés utilisés pour caractériser et classifier les instances de classes de CIDOC CRM. La classe la plus haute dans la hiérarchie de subsomption, `E1 CRM Entity` est le domaine de la propriété `P2 has type`, dont le codomaine est `E55 Type`. Ainsi, chaque classe de CIDOC CRM (à l'exception de `E59 Primitive Value`), hérite de la propriété `P2 has type`, ce qui fournit un mécanisme général pour spécialiser la classification des instances dans un graphe de connaissances utilisant CIDOC CRM à n'importe quel niveau de détail, en établissant un lien avec des sources externes, thésaurus ou ontologies. Pour classifier ainsi, il est possible d'implémenter le concept soit comme une *sous-classe* étendant le système de classes de CIDOC CRM, soit comme une *instance* de `E55 Type`. Selon les principes de construction de CIDOC CRM, une nouvelle sous-classe ne doit être créée que si le concept est suffisamment stable et associé à des propriétés supplémentaires explicitement modélisées qui lui sont propres. Sinon, une instance de `E55 Type` doit être choisie. Ce traitement cohérent des connaissances de nature terminologique renforce la capacité de CIDOC CRM à *servir de pivot d'intégration* de connaissances relevant du patrimoine culturel. En plus d'être une interface vers des thésaurus et des systèmes de classification externes, `E55 Type` est une sous-classe de `E28 Conceptual Object`. `E55 Type` et ses sous-classes héritent donc également de toutes les propriétés de cette super-classe. L'une des réflexions en cours au sein du SIG CRM porte encore sur les terminologies dans le domaine du patrimoine culturel. Une note récente de Martin Doerr précise ceci : les classes sans propriété sont modélisées comme des instances de `E55 Type`, c'est-à-dire comme des *données*. Ces données définissent un vocabulaire, sachant que les vocabulaires non standardisés sont un outil important de la recherche dans toutes les sciences et les humanités. A des fins d'interopérabilité le SIG CRM recommandera cependant dans un document distinct de la définition de CIDOC CRM certains termes, mais uniquement ceux considérés comme importants pour certaines distinctions ontologiques précises, et suffisamment univoques pour être fixés

16. `https://pro.europeana.eu/page/europeana-aat`
17. `https://ariadne-infrastructure.eu/Portal/`
18. Periodo (`perio.do`) est utilisé pour l'axe Quand et Geonames (`geonames.org`) pour l'axe Où.

comme standards. Ces termes pourront être liés ou intégrés en tant que termes plus larges ou plus étroits dans les vocabulaires utilisés par l'utilisateur, d'une manière compatible avec la signification des classes de CIDOC CRM. En outre, le SIG CRM recommandera l'utilisation de certains vocabulaires standardisés dans les cas où il existe une pratique internationale, comme pour les unités de mesure, les codes de pays, etc. Lors de la création des graphes de connaissances de OpenArchaeo, évoqués en introduction, le principe de l'utilisation de la classe `E55 Type` a été appliqué pour articuler l'utilisation de l'ontologie CIDOC CRM avec celle du thésaurus PACTOLS. Pour attribuer des labels aux entités, la question s'est posée d'utiliser `rdfs:label` ou `skos:label` et le choix suivant a été fait par les archéologues : s'il s'agit d'une entité intermédiaire (présente dans le graphe de connaissances juste pour respecter CIDOC CRM) alors un `rdfs:label` peut lui être associé si besoin, mais s'il s'agit d'une entité dont le label est un terme alors un `skos:prefLabel` lui est attaché, même si cette entité n'est pas (encore) un concept d'un thésaurus [19]. Le but étant de *marquer une intention* d'utiliser ce label comme un terme (en attendant qu'il soit dans PACTOLS).

4 Détection de termes dans un graphe de connaissances

4.1 Indices

Nous avons tenté de suivre quatre indices pour la détection des termes, résumés dans la table 1, avec trois principaux critères de comparaison. D'abord, l'objectif est d'avoir une approche avec peu de faux positifs (précision élevée) et couvrant un maximum de termes (rappel élevé). Ensuite, une approche est d'autant plus intéressante qu'elle requiert peu de connaissances d'autres ressources a priori (prérequis faible).

Classe répertoriée Parfois, les concepts dénotant des termes sont explicitement déclarés dans le graphe analysé comme instances d'une classe qui est connue pour représenter des ressources terminologiques, comme `skos:Concept`, `skos:Collection`, `crm:E55 Type` pour CIDOC CRM ou `ontolex:LexicalConcept` pour OntoLex-lemon et Termlex. En effet pour les graphes de connaissances utilisant CIDOC CRM et suivant ses recommandations (cf. section 3.2), les instances de `crm:E55 Type` (et de ses éventuelles sous-classes) doivent être des concepts dénotant des termes. Cet indice est sûr. Rien que la classe `skos:Concept` permet(trait) de couvrir de nombreux termes de par son usage fréquent. Hélas, le graphe de connaissances interrogé ne contient pas souvent cette déclaration, présente uniquement dans le thésaurus (c'est notamment le cas pour OpenArchaeo), même si cela peut arriver (Cultura Italia contient les définitions des classes, propriétés et concepts SKOS). Cet indice demande de connaître l'ontologie de référence du domaine (par exemple CIDOC CRM pour savoir que la classe `E55 Type` a pour instances des concepts de terminologies).

Propriété répertoriée Les sujets de certaines propriétés ont une forte chance d'être des concepts dénotant des termes. Nous avons d'abord pensé que c'est le cas de `skos:prefLabel`, on pourrait même espérer que ce soit un indice universel de la même manière que `rdfs:type` identifie une instance de classe. Il n'en est rien en pratique car, même si cette propriété est effectivement largement adoptée, aucune convention n'encadre son usage : sa définition ne contraint pas son domaine et il n'existe pas de bonnes pratiques clairement énoncées à notre connaissance. De ce fait la précision de cet indice avec cette propriété est faible. Le rappel est potentiellement élevé toutefois il peut arriver que, plutôt qu'un `skos:prefLabel`, ce soit un `skosxl:prefLabel` qui soit utilisé, lorsque c'est une entité lexicale qui est associée au concept, plutôt qu'une chaîne de caractères. D'autres propriétés relevant d'une description de terminologie peuvent être testées, comme d'autres propriétés de SKOS, celles de SKOS-XL, ou `ontolex:isEvokedBy` pour OntoLex-lemon et Termlex, ou les propriétés `P127 has broader term` et `P150 defines typical parts of` pour le CIDOC CRM. Mais, comme pour les classes répertoriées, ces propriétés sont rarement présentes dans le graphe analysé, mais plutôt dans le graphe où est définie la terminologie.

Préfixe répertorié Une méthode relativement naïve est de détecter les URIs correspondant à des thésauri connus. Par exemple, toutes les URIs débutant par `https://ark.frantiq.fr/` appartiennent au thésaurus PACTOLS [20]. La précision de cette approche est très élevée, mais sa couverture dépend du fait que le graphe de connaissances ciblé utilise des thésauri connus ou pas. Cette approche ne permet pas la découverte de nouveaux thésauri et elle est donc peu adaptée aux graphes de connaissances nouveaux.

Universaux filtrés En métaphysique, les universaux sont « des termes généraux qui semblent désigner ce qui est commun entre diverses choses » [11], ces termes étant utilisés pour désigner et comprendre les caractéristiques communes des *entités particulières* [15]. En ce sens, les concepts dénotant des termes sont donc des *entités universelles* au même titre que les classes ou les propriétés [21]. Nous avons cherché à détecter l'ensemble des entités universelles d'un graphe de connaissance, puis d'en retirer les classes et les propriétés. Notre méthode de détection repose sur l'idée clé qu'une *entité universelle* n'est jamais sujet d'une assertion dans laquelle une *entité particulière* est objet : une entité universelle définit une entité particulière puisqu'elle représente une caractéristique commune à

19. La propriété `skos:prefLabel` n'a pas de domaine défini et peut s'appliquer à tout type d'objet : `https://www.w3.org/TR/2009/REC-skos-reference-20090818/#L1541`

20. Il faut évidemment que tous les concepts du thésaurus partagent le même préfixe et que ce préfixe désigne uniquement des concepts dénotant des termes. Cela pose problème pour SNOMED CT.

21. Les concepteurs d'ontologies et de terminologies du Web sémantique se situeraient côté anti-réalistes, pour qui il n'existe dans la réalité que des choses particulières (ABox), les universaux se trouvant soit dans le langage (nominalistes : terminologies), soit dans l'esprit (conceptualistes ou idéalistes : TBox).

Indice	Description	Précision	Rappel	Prérequis
Classe répertoriée	Instance d'une classe dédiée à la terminologie comme `skos:Concept`, `crm:E55_Type`, etc.	très élevée	variable	élevé
Propriété répertoriée	Sujet d'une propriété en principe dédiée à la terminologie comme `skos:prefLabel`, etc.	faible	variable	élevé
Préfixe répertorié	Préfixe d'URI correspondant à un thésaurus répertorié, comme PACTOLS ou AAT du Getty	très élevée	élevé	très élevé
Universaux filtrés	Entité décrite uniquement par des littéraux ou des universaux, qui n'est ni une classe, ni une propriété	variable	élevé	faible

TABLE 1 – Quatre indices pour la détection des termes.

des entités particulières, mais elle n'est pas elle-même définie par une entité particulière. Nous recherchons donc des entités qui sont des sujets de triplets, mais pas de triplet dont l'objet est une instance de classe (entité particulière). Cela se traduit par la requête SPARQL suivante :

```
1  SELECT DISTINCT ?s
2  WHERE {
3    ?s ?p ?o .
4    FILTER NOT EXISTS {
5      ?s ?other_p ?other_o .
6      ?other_o a ?co .
7      FILTER(STR(?co)!="http://.../skos/core#Concept"
8      && !STRSTARTS(STR(?other_p),"http://.../skos"))
9    } .
10   FILTER (!ISBLANK(?s)) .
11   FILTER NOT EXISTS {?another_s ?s ?another_o} .
12   FILTER NOT EXISTS {?instance a ?s}
13 }
```

Cette requête recherche les entités universelles en prenant toutes les entités qui sont le sujet `?s` d'au moins une assertion (ligne 3), qui ne réfèrent à aucune entité particulière (ici, instance d'une classe avec la ligne 5 et 6) et qui ne sont pas des *blank nodes* (ligne 10). De plus, on ne souhaite pas éliminer les cas où la propriété `?other_p` appartient à l'ontologie `SKOS` (ligne 8) afin de respecter l'indice « Propriété répertoriée », ni ceux où la classe `?co` est `skos:Concept` (ligne 7), pour respecter l'indice « Classe répertoriée ». Enfin, les lignes 11 et 12 filtrent les termes en empêchant respectivement que l'entité `?s` soit une propriété et une classe. La force de cette proposition est de nécessiter peu de connaissances prérequises sur les graphes à analyser [22], ce qui est particulièrement adapté pour la détection de termes dans des graphes de connaissances nouveaux. Son rappel risque d'être plus élevé que les autres approches mais elle risque de détecter des faux positifs, à savoir : des entités particulières reliées à aucune autre entité particulière d'une part, et d'autre part des entités universelles qui ne seraient ni des concepts dénotant des termes, ni des classes/propriétés.

4.2 Expérimentations

Nous reportons dans cette section nos premières expérimentations correspondant à l'application des quatre méthodes de détection sur dix graphes de connaissances disponibles via des points d'accès publics : ADS [23] qui regroupe les connaissances en archéologie du Royaume-Uni, Cultura Italia [24] qui contient des connaissances sur des collections de musées et galeries italiennes, et 8 graphes de OpenArchaeo [25], portail sémantique d'une communauté d'archéologues regroupée au sein du consortium MASA [26] de la TGIR Huma-Num. Le tableau 2 donne quelques informations statistiques (nombre d'entités et d'assertions) et il présente les résultats obtenus, avec le nombre de concepts dénotant des termes découverts pour chaque méthode. Ces résultats sont obtenus en testant les classes, propriétés et thésauri cités dans la colonne Description du tableau 1. A noter qu'un usage immodéré de la propriété `skos:prefLabel` rend illusoire l'utilisation de la méthode « Propriété répertoriée » pour les graphes de OpenArchaeo. De la même façon, il y a dans Cultura Italia 866 instances distinctes de `skos:Concept` et 358 492 instances distinctes de `E55 Type` : leur analyse montre que 1 600 d'entre eux sont bien des termes d'un thesaurus, lequel est contenu dans le graphe de connaissances, mais que tout le reste correspond à des URIs décrivant des *valeurs littérales de dimensions*, comme diamètre, hauteur, etc. et de combinaisons de telles dimensions, comme par exemple `ci:format/cm-diametro-29-peso-20-6` [27]. Là encore nous pouvons conclure à un usage immodéré, voire injustifié de la classe `E55 Type`.

Les concepts dénotant des termes trouvés avec la méthode des universaux filtrés sont trouvés avec les méthodes « Classe répertoriée » ou « Propriété répertoriée » en ce qui concerne les graphes ADS et Cultura Italia. Pour les graphes de OpenArchaeo, les concepts dénotant des termes retrouvés appartiennent aux thésaurus PACTOLS et Getty. Pour ADS, les concepts dénotant des termes retrouvés appartiennent à différents thésauri (archaide, romanamphorae, etc.) qui nous étaient inconnus. Dans plusieurs graphes de OpenArchaeo, cette méthode ne donne rien parce que les concepteurs ont associé aux concepts du thésaurus PACTOLS un autre terme que dans le thesaurus en utilisant la propriété `skosxl:altLabel`. Dans un objectif d'interopérabilité, ils utilisent systématiquement un concept du thésaurus PACTOLS pour certaines propriétés du graphe, mais lorsque dans le jeu de données d'origine, fourni par

22. On peut en ajouter davantage au niveau des lignes 7 et 8.
23. `http://data.archaeologydataservice.ac.uk/query/`
24. `http://dati.culturaitalia.it/sparql`
25. `http://openarchaeo.huma-num.fr/federation/sparql`
26. `https://masa.hypotheses.org/`
27. ci <`http://dati.culturaitalia.it/resource/`>

Graphe	Entités	Assertions	Classe répertoriée	Propriété répertoriée	Préfixe répertorié	Universaux filtrés
ADS	214 155	1 547 192	1863	1588	-	1251
Cultura Italia	9 951 821	41 901 551	359358	0	981	749
arsol	105 054	669 099	75	-	150	0
chronique	98 837	557 724	114	-	118	1
epicherchell	706	3 945	18	-	32	0
iceramm	7 456	44 538	42	-	581	0
kition-pervolia	6 490	32 714	93	-	114	94
outagr	30 313	115 462	400	-	584	414
rita	12 366	76 885	518	-	479	0
solidar	9 848	48 460	94	-	159	127

TABLE 2 – Détection des concepts termes, en utilisant les différentes méthodes.

	Faux positifs				
Graphe	Part.	Prop.	Classes	Types	Autre
ADS	539	0	0		0
Cultura Italia	0	2	4	59	0
aerba	0	0	0	0	0
arsol	0	0	0	0	0
chronique	0	0	0	0	1
epicherchell	0	0	0	0	1
iceramm	0	0	0	0	1
kition-pervolia	2140	0	0	0	1
outagr	14121	0	0	0	1
rita	0	0	0	0	1
solidar	1550	0	0	0	0

TABLE 3 – Analyse des FP avec le filtrage des universaux

les chercheurs, ce sont d'autres mots qui sont utilisés ils les ajoutent avec `skosxl:altLabel`. Ils ont choisi d'utiliser la propriété SKOS-XL pour dénoter que ces mots ne viennent pas de PACTOLS mais de leur jeu de données original. Le problème pour notre méthode des universaux filtrés est que cette propriété prend pour objet une instance de la classe `skosxl:Label` et que les créateurs déclarent dans le graphe que ces objets sont de cette classe. De plus, aussi bien ces objets que les concepts de PACTOLS ou AAT Getty sont systématiquement déclarés instances de `owl:NamedIndividual` par un effet de bord d'une utilisation du logiciel Protégé, que les créateurs pensent sans conséquence pour leur exploitabilité. Pourtant, déclarer que ces entités universelles que sont les concepts de thésaurus sont des « named individual » modifie significativement le regard d'un point de vue philosophique. La méthode des universaux filtrés est ainsi tributaire d'aléas dans le processus de création des graphes.

Nous analysons maintenant les résultats retournés par cette méthode qui ne sont pas des concepts dénotant des termes. Le tableau 3 détaille ces faux positifs pour la méthode fondée sur le filtrage des universaux. La précision de cette approche, vérifiée en examinant tous les résultats obtenus, est très variable : 69,9% pour ADS, 92,0% pour Cultura Italia et seulement 3,4% sur les graphes de OpenArchaeo.

Pour ADS, les faux positifs correspondent à des entités particulières, URL de documents. Pour Cultura Italia, certaines propriétés non-utilisées et classes non-instanciées ont été retournées étant donné que ce graphe comporte toutes les définitions des classes et propriétés des ontologies utilisées. Les autres faux positifs correspondent à des typages de littéraux (entiers, chaîne de caractères,...), qui peuvent aussi être considérés comme des entités universelles et pourraient être retirées en complétant notre requête SPARQL. Pour les graphes de OpenArchaeo, dans les faux positifs, seulement 3 réponses correspondent à des entités universelles (noms de graphes) et le reste correspond à des entités particulières réparties en deux catégories. La première représente des entités X, qui existent dans le jeu de données d'origine mais ne sont présentes dans aucune page web où ce jeu de données est présenté (autrement il s'agit de l'URL de la page web qui est utilisé, terminé par #X car une même page web peut montrer plusieurs entités X). La seconde catégorie contient des URLs vers d'autres graphes, par exemple vers des entités de GeoNames. Pour ces entités-là, notre requête aurait fonctionné si elle avait été appliquée conjointement sur les graphes de OpenArchaeo et sur ceux dont sont issues les entités sélectionnées.

Le tableau 2 montre donc qu'aucune des méthodes imaginées ne fonctionne parfaitement. Les propriétés ou classes répertoriées peuvent servir mais leur utilisation est délicate si les graphes de connaissances les utilisent pour autre chose que des éléments de terminologies ou de thésauri. La méthode des universaux filtrés est également dépendante de choix de conception des graphes, qui restent très libres en l'absence de conventions bâties sur un consensus. Il n'est pas toujours possible d'identifier les concepts dénotant des termes via le préfixe de leur URI, mais lorsque cela est possible cette option s'avère la meilleure. Les autres méthodes peuvent permettre d'identifier des préfixes de thésauri ou de terminologies utilisés dans les graphes.

5 Conclusion

Dans cet article, nous nous sommes intéressés à l'usage de terminologies dans les graphes de connaissances du Web, en focalisant sur le domaine du patrimoine culturel où elles viennent compléter l'ontologie CIDOC CRM. Après avoir rappelé les raisons d'exister des terminologies, nous avons noté qu'elles se manifestent dans le Web sémantique soit

comme des ontologies, soit comme des thésauri, définis soit avec SKOS, soit avec d'autres vocabulaires. Dans le premier cas (ontologies) elles apparaissent dans les profils de graphes de connaissances qui montrent les classes et les propriétés utilisées, ce qui répond à notre motivation initiale, qui est de montrer dans le profil d'un graphe de connaissances des éléments de terminologies qu'il contient. Dans le deuxième cas (thésauri), il est compliqué de les détecter car pour l'heure dans le Web sémantique la définition et l'usage des thésauri ne sont pas cadrés par des définitions ou par des usages standards précisés dans des guides de bonnes pratiques, à l'image de la définition et l'usage d'ontologies. Nous avons testé plusieurs méthodes pour identifier des éléments de ressources terminologiques dans des graphes de connaissances, avec des résultats préliminaires sur trois points d'accès SPARQL offrant des graphes liés au patrimoine culturel. Dans le cadre du patrimoine culturel, pour des graphes conçus avec le CIDOC CRM en suivant ses principes pour l'utilisation conjointe de terminologies, il est possible de trouver des termes avec l'une ou l'autre des trois premières méthodes, mais ce n'est pas généralisable. Nous avons proposé et testé une méthode plus agnostique, qui échoue toutefois dans son état actuel à distinguer les particuliers des universels dans la plupart des graphes testés. Une partie des raisons de ces échecs révèle des anomalies de conception et pourrait être résolue en modifiant les graphes, une autre pourrait l'être en raffinant la méthode. Ce sont-là les deux perspectives que nous allons explorer.

Remerciements

Ce travail est financé par l'ANR-18-CE38-0009 SESAMES. Les auteurs remercient les relecteurs anonymes et Thomas Francart et Yannick Duthé pour leurs critiques et suggestions.

Références

[1] Fabien Amarger, Catherine Roussey, Jean-Pierre Chanet, Olivier Haemmerlé, and Nathalie Hernandez. Etat de l'art : Extraction d'information à partir de thésaurus pour générer une ontologie. In *INFORSID*, pages 29–44, 2013.

[2] Chryssoula Bekiari, George Bruseker, Martin Doerr, Christian-Emil Ore, Stephen Stead, and Athanasios Velios. Definition of the CIDOC Conceptual Reference Model. last official version : 7.1.1. *ICOM/CIDOC Documentation Standards Group. CIDOC CRM SIG*, 2021.

[3] Lamine Diop, Arnaud Giacometti, Béatrice Markhoff, and Arnaud Soulet. TTProfiler : Computing Types and Terms Profiles of Assertional Knowledge Graphs. In *Proceedings of the Semantic Web and Ontology Design for Cultural Heritage workshop co-located with (BOSK 2021)*, volume 2949 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2021.

[4] FRANTIQ. Le thésaurus pactols. `https://www.frantiq.fr/pactols/le-thesaurus/`, 2020. Accessed on 2022-28-02.

[5] Fabien Gandon, Catherine Faron-Zucker, and Olivier Corby. *Le Web sémantique*. Dunod, Paris, France, 2012.

[6] François Goasdoué, Pawel Guzewicz, and Ioana Manolescu. RDF graph summarization for first-sight structure discovery. *VLDB J.*, 29(5) :1191–1218, 2020.

[7] Aidan Hogan, Eva Blomqvist, Michael Cochez, Claudia d'Amato, Gerard de Melo, Claudio Gutiérrez, Sabrina Kirrane, José Emilio Labra Gayo, Roberto Navigli, Sebastian Neumaier, Axel-Cyrille Ngonga Ngomo, Axel Polleres, Sabbir M. Rashid, Anisa Rula, Lukas Schmelzeisen, Juan F. Sequeda, Steffen Staab, and Antoine Zimmermann. Knowledge graphs. *ACM Comput. Surv.*, 54(4) :71 :1–71 :37, 2021.

[8] D. Kless, L. Jansen, and S. Milton. A content-focused method for re-engineering thesauri into semantically adequate ontologies using owl. *Semantic Web*, 7(5) :543–576, 2016.

[9] D. Kless, S. Milton, E. Kazmierczak, and J. Lindenthal. Thesaurus and ontology structure : Formal and pragmatic differences and similarities. *Journal of the Association for information science and technology*, 66(7) :1348–1366, 2015.

[10] Marion Lamé, Perrine Pittet, Federico Ponchio, Béatrice Markhoff, and Emilio M. Sanfilippo. Heterotoki : non-structured and heterogeneous terminology alignment for digital humanities data producers. In *Workshop on Open Data and Ontologies for Cultural Heritage co-located with CAiSE, ODOCH@CAiSE 2019*, volume 2375 of *CEUR Workshop Proceedings*, pages 37–48. CEUR-WS.org, 2019.

[11] Bruno Langlet. Universaux (GP), dans Maxime Kristanek (Dir.), l'Encyclopédie philosophique. `https://encyclo-philo.fr/universaux-gp`, 2019. Consulté le 15/03/2022.

[12] Bernadette Farias Lóscio, Caroline Burle, and Newton Calegari. Data on the web best practices (w3c recommendation 31 january 2017) : Best practice 15. `https://www.w3.org/TR/dwbp/`, 2017. Accessed on 2022-28-02.

[13] Emmanuelle Perrin. Thésaurus et interopérabilité des données archéologiques : le projet hyperthesau. *Humanités numériques [en ligne]*, 4, 2021.

[14] Christophe Roche and Maria Papadopoulou. Terminology and ontology for digital humanities : The case of ancient greek dress. *Humanités numériques*, 2, 2020.

[15] Bertrand Russell. *On the relations of universals and particulars*. Lulu Press, Inc, 2015.

[16] Blerina Spahiu, Riccardo Porrini, Matteo Palmonari, Anisa Rula, and Andrea Maurino. ABSTAT : Ontology-Driven Linked Data Summaries with Pattern Minimalization. In *The Semantic Web - ESWC 2016 Satellite Events, Revised Selected Papers*, pages 381–395. Springer, 2016.