



MANON OVIDE, BÉATRICE MARKHOFF, ARNAUD SOULET

De l'importance des termes dans les graphes de connaissances et comment les trouver
Volume 6, n° 1-2 (2025), p. 131-155.

<https://doi.org/10.5802/roia.96>

© Les auteurs, 2025.



Cet article est diffusé sous la licence
CREATIVE COMMONS ATTRIBUTION 4.0 INTERNATIONAL LICENSE.
<http://creativecommons.org/licenses/by/4.0/>



*La Revue Ouverte d'Intelligence Artificielle est membre du
Centre Mersenne pour l'édition scientifique ouverte*
www.centre-mersenne.org
e-ISSN : 2967-9672

De l'importance des termes dans les graphes de connaissances et comment les trouver

Manon Ovide^a, Béatrice Markhoff^b, Arnaud Soulet^a

^a Université de Tours, UR 6300 LIFAT (France)

E-mail : manon.ovide@univ-tours.fr, arnaud.soulet@univ-tours.fr

^b Université de Tours, UMR 7324 CITERES – LAT (France)

E-mail : beatrice.markhoff@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 informent sur ce contenu au moins autant que les classes et les propriétés : les termes de vocabulaires contrôlés, des mots associés à des concepts définis au sein de thésaurus. À partir de 23 points d'accès SPARQL, nous présentons différentes techniques pour retrouver automatiquement des termes qui ne sont pas des noms de classes. Les résultats obtenus par chaque technique nous permettent de proposer une méthodologie pour trouver les termes présents dans un graphe de connaissances. Les résultats des expérimentations et les requêtes support de la méthodologie sont accessibles en ligne.

MOTS-CLÉS. — CIDOC CRM, graphe de connaissances, ontologie, SPARQL, terme, terminologie, thésaurus.

1. INTRODUCTION

Les termes sont particulièrement porteurs de connaissances : par exemple en archéologie, 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 « correctement » les objets découverts, leurs fonctions, leur forme, leur constitution (matériaux) et toutes leurs caractéristiques. Il existe une tension récurrente entre la standardisation et la liberté de choix et de précision des termes utilisés, choix et précision qui relèvent de l'activité savante. Partager cette activité et ses conclusions suppose pourtant un consensus, au moins à un assez haut niveau d'abstraction [18]. 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 archéologique française a développé un système de gestion de thésaurus [21] et un thésaurus [10], 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 de choix et de définition des 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. Cela est fait avec une motivation double : les archéologues désirant *décrire* un objet d'étude et son étude en elle-même, les documentalistes souhaitant *indexer pour retrouver* l'information. En effet, connaître le vocabulaire utilisé pour décrire les résultats de fouille est crucial pour interroger les bases de données qui conservent ces résultats, ceci tout autant que connaître le schéma conceptuel de la base de données. Il en va de même lorsqu'on veut interroger les graphes de connaissances.

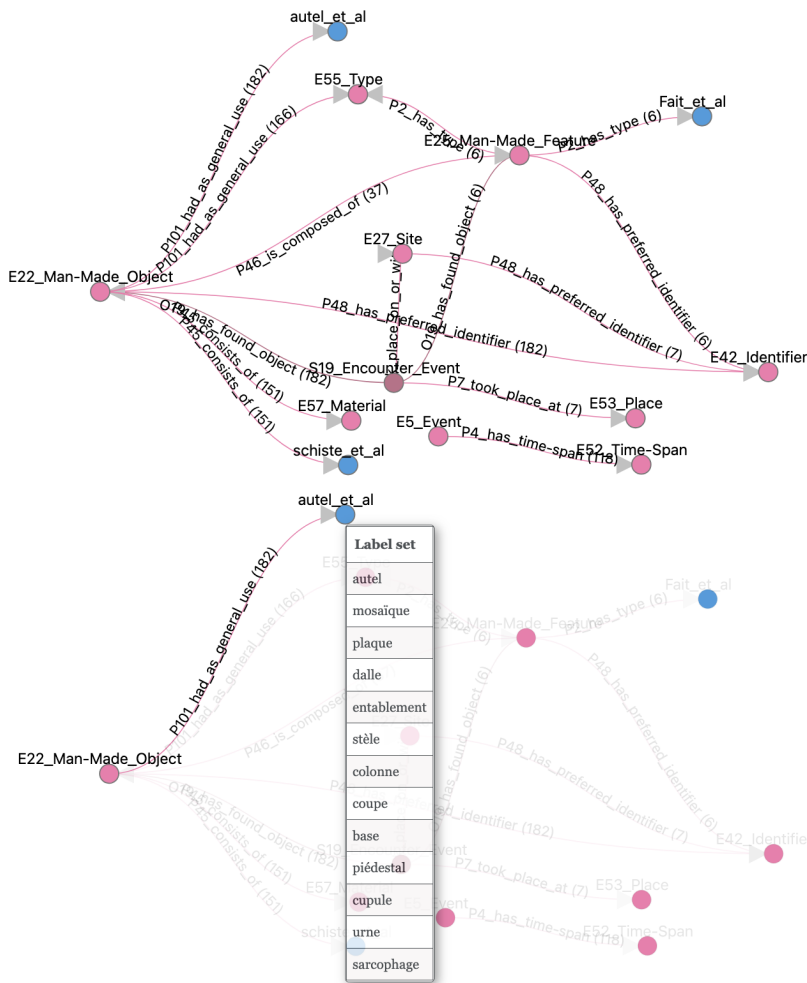


FIGURE 1.1. Profil du graphe Epicherschell de OpenArcheo.

Ainsi pour le graphe Epicherchel comme pour les autres graphes de connaissances du portail OpenArchaeo⁽¹⁾ auquel nous contribuons, certaines propriétés ont pour objet un concept décrit dans le thesaurus PACTOLS⁽²⁾, dont le terme associé informe très précisément sur le contenu du graphe. On voit en figure 1.1 que c'est le cas des propriétés *P2 has type*⁽³⁾, *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* qui sont des concepts, représentés par le noeud *autel et al.*. Il est très utile de montrer les labels de ces concepts, ou termes : savoir qu'il y a des milliers d'instances de *E22 Man-Made Object* dans le graphe indique juste que ce graphe contient des informations sur des objets du patrimoine culturel, alors que voir les termes rassemblés dans le noeud *autel et al.* permet de comprendre ce que sont les objets décrits dans le graphe. Ces éléments sont essentiels pour requêter le graphe : ainsi, si on veut requêter des autels dans le graphe Epicherchell, on devra utiliser le terme de label « autel » plutôt que rechercher des instances d'une classe Autel (inexistante).

Plus généralement, utiliser des termes pour encoder les données et les métadonnées fait partie des *bonnes pratiques dans la production* de graphes de connaissances [9] : les bénéfices sont la réutilisabilité, la facilité de traitement (requêtes), la compréhension, la confiance et l'interopérabilité. Les termes doivent appartenir à des vocabulaires, de préférence partagés, voire standardisés. Dans ces vocabulaires, les termes sont associés à des URIs, décrits de façon non ambiguë et qui peuvent avoir des labels (termes) dans différentes langues. Ces URIs peuvent dénoter des classes, ou pas.

Du côté de la *consommation* d'un graphe de connaissances, il est bien connu qu'il est très difficile pour tout utilisateur d'un graphe de connaissances qu'il n'a pas lui-même créé d'en découvrir le contenu. Pour cela, diverses propositions ont été faites pour résumer [5, 13] ou profiler [24] un graphe de connaissances. Ces profils ou résumés montrent les classes et les propriétés d'ontologies utilisées dans le graphe. C'est le cas de nos précédents travaux [7, 8], qui ont pour originalité de montrer également des termes de thésaurus utilisés. Par exemple, pour le graphe de connaissances Epicherchel, qui contient les objets sur lesquels ont été trouvées des inscriptions antiques à Césarée de Maurétanie, le profil en figure 1.1⁽⁴⁾, généré par [8], montre que le graphe contient des instances des classes *E27 Site* et *E22 Man-Made Object*, entre autres. Il permet également de savoir qu'il y a 182 triplets (sujet, prédicat, objet) où le sujet est de la classe *E22 Man-Made Object*, l'objet de la classe *E55 Type* et le prédicat est la propriété *P101 had as general use*. Mais alors que dans [8] seules 3 heuristiques sont utilisées pour générer des profils montrant des termes, ici l'accent est mis sur leur détection systématique, approfondissant ainsi cette question : comment trouver des termes qui ne sont pas des noms de classe ou de propriété, mais des éléments de thésaurus. *La problématique que nous abordons* est la suivante : alors que ces termes sont présents

⁽¹⁾<http://openarchaeo.huma-num.fr/explorateur/home>

⁽²⁾<https://www.frantiqu.fr/pactols/le-thesaurus/>

⁽³⁾Les classes et propriétés appartiennent au CIDOC CRM : <https://cidoc-crm.org>

⁽⁴⁾Visible en ligne ici : <https://kgsumviz.univ-tours.fr/Home.php>

dans de nombreux graphes de connaissances⁽⁵⁾, il n'est pas toujours simple de les caractériser car on ne sait pas a priori quels thésaurus sont utilisés. Donc on ne peut pas toujours les trouver automatiquement pour les montrer dans un profil. Les contributions de cet article sont d'enrichir l'étude préliminaire menée dans [20] grâce à l'élargissement du corpus de graphes étudiés, la validation systématique des hypothèses qu'elle proposait (l'une des requêtes qui y étaient proposées a été abandonnée) et la proposition d'une méthodologie pour trouver les termes de thésaurus.

Dans la suite de cet article, la notion de terminologie et ses représentations (les thésaurus – dont la motivation est d'indexer pour retrouver – et les ontologies – dont la motivation est de modéliser des connaissances –), sont rappelées en section 2. Nous donnons en section 3 un aperçu de l'importance et de la variété des ressources terminologiques dans le Web, en nous intéressant aux relations que les ontologies et les thésaurus peuvent entretenir. Nous faisons un focus sur l'ontologie CIDOC CRM [2] pour le patrimoine culturel, mais aussi sur deux autres ressources, emblématiques du Web sémantique : Schema.org et Wikidata. Les deux sections suivantes ont pour objet de déterminer *comment détecter des termes* dans un graphe de connaissances. Car autant la propriété `rdf:type` dénote une instance de classe et la deuxième position dans un triplet dénote une instance de propriété, autant détecter un terme de thésaurus ne repose sur aucun élément de syntaxe aussi simple. Savoir qu'une ressource terminologique est utilisée est une étape déterminante, que nous décrivons dans la section 4. Dans la section 5, nous donnons d'autres formes de syntaxe pouvant dénoter des termes et les requêtes SPARQL correspondantes. Dans ces deux sections, nous présentons des résultats d'expérimentations sur 23 points d'accès SPARQL, résultats sur lesquels repose la méthodologie de détection de termes que nous recommandons. Enfin, nous concluons en section 6.

2. RÔLE ET REPRÉSENTATION DES TERMES

Les termes, ou *mots associés à des concepts définis*, sont des éléments de terminologie. Ils peuvent relever de thésaurus ou d'ontologies. Nous commençons par resituer ces notions les unes par rapport aux autres.

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 façon 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)

⁽⁵⁾ Graphe de connaissances au sens défini dans [15] qui, de façon intéressante, ne parle pas de l'existence de ces termes.

qui ont inspiré l'ontologie OntoLex-lemon⁽⁶⁾ (représentation des propriétés morpho-syntaxiques des entrées lexicales et de leur sens) et son extension pour la terminologie en cours de définition, Termlex⁽⁷⁾, 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 de la classe *LexicalSense*, qui peut faire le lien entre un élément de lexique et une ontologie dans laquelle le concept correspondant est décrit⁽⁸⁾. La question de ce lien entre forme lexicale et concept pour la description de termes a également mené à la notion de système onto-terminologique [22].

Cependant en général dans le Web, certaines terminologies sont réalisées juste sous forme d'ontologie, comme le Dublin Core⁽⁹⁾ où les termes sont des classes, et d'autres sont réalisées juste avec SKOS ou SKOS-XL⁽¹⁰⁾, comme le AAT du Getty⁽¹¹⁾. SKOS a pourtant une expressivité limitée pour décrire la sémantique d'un terme, puisque c'est une ontologie de thésaurus, taxonomies, schémas de classification ou systèmes de vedettes-matières, conçus à des fins d'indexation et de recherche d'information.

2.2. INDEXATION ET RECHERCHE D'INFORMATION : LES THÉSAURUS

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 thésaurus 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*. Tandis que l'objectif principal des thésaurus 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 subsumption, la relation de partition, parfois aussi la relation d'instanciation, fusionnées en une relation hiérarchique unique 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

⁽⁶⁾<https://www.w3.org/2016/05/ontolex/>

⁽⁷⁾<https://www.w3.org/community/ontolex/wiki/Terminology>

⁽⁸⁾Voir <http://anr-sesames.map.cnrs.fr/onto/shart/index.html> pour un exemple d'utilisation

⁽⁹⁾Voir <https://www.dublincore.org/specifications/dublin-core/dcmi-terms/>

⁽¹⁰⁾<https://www.w3.org/TR/skos-reference/>

⁽¹¹⁾Art and Architecture Thesaurus : <https://www.getty.edu/research/tools/vocabularies/aat/>

de B devraient être retournés ». Les grands thésaurus sont organisés en facettes, qui regroupent des hiérarchies de concepts pour faciliter la recherche d'information. Comme le note [17], la structure des thésaurus ne peut pas être la base de raisonnements plus généraux que le simple accès par indexation, contrairement aux ontologies.

2.3. MODÉLISATION DE CONNAISSANCES : LES ONTOLOGIES

Il n'est pas question ici de définir précisément ce qu'est une ontologie dans le Web comme le font les auteurs de [17], rappelons simplement que c'est un modèle 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 subsumption 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 subsumption 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 [12] 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 subsumption, 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 avec un thésaurus il faut développer un système ad hoc pour réaliser cette généralisation de la requête, il n'y a pas un raisonneur générique pour le faire. La bonne pratique précédente est suivie dans le domaine des sciences de la terre et du vivant, où les terminologies sont représentées en OWL2 EL, c'est-à-dire sous forme d'ontologies qui contiennent un grand nombre de classes et de propriétés. Cela peut se vérifier dans des ontologies référencées dans BioPortal⁽¹²⁾.

Il existe de nombreuses ressources de connaissances structurées sous la forme de thésaurus en SKOS produites par les bibliothèques. Comme une structuration sous forme d'ontologie offre pour avantages le raisonnement automatique et la précision dans la description des relations entre concepts, la tâche de transformer en ontologie les connaissances contenues dans un thésaurus pour les exploiter automatiquement avec un raisonneur est abordée de longue date [1, 16]. Cette tâche n'est pas simple, notamment du fait que la relation `broader-narrower` d'un thésaurus peut être aussi bien une relation de subsumption 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.

Plutôt que transcrire une forme dans une autre, il est aussi possible de les articuler dans une utilisation conjointe. Dans le domaine du patrimoine culturel, la communauté

⁽¹²⁾<https://bioportal.bioontology.org/>

qui définit et maintient l'ontologie CIDOC CRM a une politique claire pour l'usage de descriptions terminologiques conjointement avec l'ontologie. Dans un tout autre registre, une articulation comparable est permise dans l'ontologie Schema.org [26], à l'origine d'un très gros volume de données sémantiques dans le Web. Enfin, les conventions adoptées par les concepteurs de la plus grande ressource sémantique du Web, Wikidata [26], conduisent à un rapport original entre thésaurus et ontologie : les deux formes y coexistent, portées par les mêmes entités, c'est *l'utilisation* de ces entités qui en fait soit des termes de thésaurus, soit des classes. Nous présentons dans la section suivante ces différentes approches.

3. DES THÉSAURUS ET DES ONTOLOGIES DU WEB

3.1. RESSOURCES TERMINOLOGIQUES DANS LE WEB

On trouve des ressources terminologiques sous la forme de thésaurus 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 thésaurus sont plus massivement réutilisés que des ontologies conçues pour représenter un domaine de connaissance, dans la mesure où le consensus nécessaire à leur réutilisation porte sur les mots (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 plus des ressources terminologiques que des conceptualisations d'un domaine de connaissance. Il existe de nombreuses et, pour certaines, très grandes ressources terminologiques dans le Web. Par exemple dans le domaine biomédical UMLS⁽¹³⁾ rassemble des concepts de plusieurs dizaines de terminologies, MeSH⁽¹⁴⁾ (défini par la bibliothèque nationale de médecine des États-Unis) permet d'indexer les répertoires d'articles Medline et PubMed, quand SNOMED CT⁽¹⁵⁾ 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⁽¹⁶⁾ 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⁽¹⁷⁾ 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 du Getty, déjà cité également, qui est utilisé et avec lequel les autres thésaurus sont alignés, par exemple dans EUROPEANA⁽¹⁸⁾ ou dans la plateforme

⁽¹³⁾Unified Medical Language System : <https://www.nlm.nih.gov/research/umls/index.html>

⁽¹⁴⁾Medical Subject Headings : <https://www.ncbi.nlm.nih.gov/mesh>

⁽¹⁵⁾Systematized Nomenclature Of Medicine Clinical Terms : <https://www.snomed.org/>

⁽¹⁶⁾<https://www.fao.org/agrovoc/>

⁽¹⁷⁾<https://www.backbonethesaurus.eu/>

⁽¹⁸⁾<https://pro.europeana.eu/page/europeana-aat>

ARIADNE⁽¹⁹⁾. Celle-ci permet l'interrogation selon trois axes Quand-Où-Quoi : le Getty AAT est utilisé pour décrire ce qui est recherché (Quoi)⁽²⁰⁾.

À noter que pour les axes 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*, liées chacune à un référent unique, et non plus de *termes* « universaux » [19]. Du point de vue de *leur utilisation dans un graphe de connaissances*, il est fait référence à ces entités nommées comme à des éléments d'une source d'autorité, extérieure et partagée, qui servent à caractériser des individus du graphe : elles sont donc utilisées de la même façon que des termes. Il est certain que montrer des entités nommées présentes dans un graphe de connaissances est également très informatif sur son contenu, mais ce n'est pas le sujet ici.

3.2. CAS DU PATRIMOINE CULTUREL : CIDOC CRM

CIDOC CRM est une ontologie conçue pour supporter l'interopérabilité sémantique de descriptions numériques du patrimoine culturel. Elle est 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 subsumption, E1 CRM Entity, est le domaine de la propriété P2 has type, dont le co-domaine est E55 Type. Ainsi, toute instance d'une classe de CIDOC CRM peut être caractérisée via la propriété P2 has type. C'est 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. Pour illustrer ce mécanisme, la figure 3.1 donne un exemple de représentation d'une pièce de monnaie où l'on utilise la ressource terminologique Nomisma⁽²¹⁾ pour caractériser le fait que l'objet décrit est une pièce de monnaie.

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. Il est rappelé que dans le CIDOC CRM *les classes sans propriété* ne peuvent apparaître que comme instances de E55 Type, c'est-à-dire comme des *données*, et que ces données définissent un vocabulaire

⁽¹⁹⁾<https://ariadne-infrastructure.eu/Portal/>

⁽²⁰⁾Periodo (perio.do) est utilisé pour l'axe Quand et GeoNames (geonames.org) pour l'axe Où.

⁽²¹⁾nomisma.org

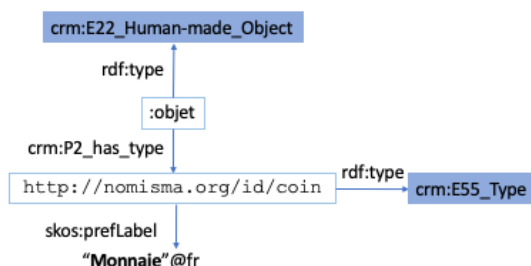


FIGURE 3.1. Exemple de représentation d'une pièce de monnaie avec CIDOC CRM.

contrôlé. Ce mécanisme est important pour permettre de définir des vocabulaires internes non standardisés, car les vocabulaires sont des outils importants de la recherche dans toutes les sciences et surtout dans les humanités. À des fins d'interopérabilité, le SIG CRM pourrait recommander l'usage de certains termes, uniquement ceux utiles pour certaines distinctions ontologiques précises et suffisamment univoques pour être fixés comme standards. Ces termes pourront être liés ou intégrés en tant que termes plus généraux (broader) ou plus précis (narrower) dans les vocabulaires propres à un graphe de connaissances. En outre, le SIG CRM recommanderait l'utilisation de certains vocabulaires standardisés dans les cas où il existe une pratique internationale (par exemple, pour les unités de mesure ou les codes de pays).

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 savant* alors un `skos:prefLabel` est utilisé, même si cette entité n'est pas encore un concept d'un thésaurus. Le but étant de *marquer une intention* d'utiliser cette entité comme un terme⁽²²⁾.

3.3. RESSOURCES MULTIDOMAINES DANS LE WEB : SCHEMA.ORG, WIKIDATA

L'étude du CIDOC CRM, pensé pour fédérer la représentation des connaissances et les terminologies dans le domaine du patrimoine culturel, fournit des indices précieux pour trouver les termes dans les graphes de connaissances de ce domaine. Il est intéressant de constater que le même mécanisme d'extension de la caractérisation des entités par des terminologies se retrouve dans Schema.org [14]. Il y a en effet une classe `schema:DefinedTermSet` permettant de déclarer tout ensemble de termes définis, système de classification, glossaire, dictionnaire ou énumération, ainsi qu'une classe

⁽²²⁾Même si 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>

`schema:DefinedTerm`, représentant tout terme associé à une définition formelle. La propriété `schema:inDefinedTermSet` permet de relier des termes à la terminologie où ils sont définis. Ces mécanismes ont été ajoutés dans l'objectif de permettre d'utiliser des terminologies existantes dans le Web, listant des types de sports, des types de comptes bancaires, des types de traitements médicaux, des types d'éléments chimiques, etc. Rappelons que Schema.org est une ontologie généraliste initiée en 2011 par les propriétaires des grands moteurs de recherche afin que les concepteurs de sites Web décrivent les données contenues dans leurs pages Web de façon à ce qu'elles soient exploitables au niveau sémantique. Depuis 2015, ses extensions sont menées par une communauté du W3C⁽²³⁾. Cette ressource est utilisée à très grande échelle et une initiative, The Web Data Commons, maintient depuis 2013 un site recensant les données RDF produites par ce biais⁽²⁴⁾ : d'après [3], il s'y trouvait en 2023 des descriptions de plus de 3 milliards d'entités provenant d'environ 13 millions de sites Web. Des auteurs tentent d'analyser les usages de cette ontologie, par exemple [6]. Mais à notre connaissance les travaux existants s'intéressent uniquement aux classes utilisées, jamais aux termes de vocabulaires contrôlés potentiellement présents dans ces graphes.

La généralité de Schema.org l'a fait adopter par les concepteurs de la ressource sémantique généraliste YAGO⁽²⁵⁾, dont la particularité est de garantir la validité des raisonnements automatiques sur des données issues des *communs de connaissances* que sont Wikipedia, DBpedia, puis Wikidata. Pour permettre ces raisonnements, depuis sa première version en 2008 l'accent a été mis sur la qualité de son schéma, en exploitant des structures ouvertes telles WorldNet, puis Schema.org. Dans sa version 4.5 [25], les auteurs rapportent que les utilisateurs de YAGO souhaiteraient disposer de plus de classes correspondant à des éléments de Wikidata, difficiles à définir uniquement à l'aide de propriétés, comme par exemple *financial regulatory agency*. Les auteurs de YAGO 4.5 extraient de Wikidata (i) une taxonomie de haut niveau (*upper taxonomy*) de classes, ayant des propriétés dont les objets relèvent également de classes de haut niveau (objectif de représentation du monde), (ii) une taxonomie inférieure (*lower taxonomy*) dont les classes fournissent des informations aux humains pour décrire des instances (objectif de caractérisation d'entités par des termes) et (iii) des instances. Les arguments mobilisés pour faire ces distinctions rejoignent le propos du présent article sur les différentes représentations des termes (par des classes d'ontologies ou par des concepts de thésaurus).

Pivot formel des centaines de Wikipédias existants dans les différentes langues, Wikidata est une énorme ressource sémantique du Web, initiée en 2013, augmentée et régulée par une communauté de contributeurs. Comme le rappellent les auteurs de [25], les choix de représentation de connaissances faits par les initiateurs de Wikidata [26] en font une exception en ce qui concerne la représentation des terminologies sous forme d'ontologies ou de thésaurus. En effet, dans Wikidata c'est la même entité qui peut être à la fois une classe ou une instance. Une entité peut apparaître comme objet de la propriété *instance of* (`wdt:P31`) ou sujet de la propriété *subclass of* (`wdt:P279`)

⁽²³⁾<https://www.w3.org/community/schemaorg/>

⁽²⁴⁾<http://webdatacommons.org/structureddata/>

⁽²⁵⁾<https://yago-knowledge.org/>

et en même temps comme sujet de `wdt:P31`. Par exemple l'entité *astronaute* (Q11631) est à la fois une instance de *profession* (Q28640) et une sous-classe de *scientifique* (Q901). Ainsi Wikidata permet d'utiliser une même entité parfois comme un élément de thésaurus, c'est-à-dire un couple concept-terme doté d'une définition et dont l'usage est de caractériser une autre entité, et parfois comme une définition en intension d'un élément du monde, dont l'usage est de représenter tout un ensemble d'instances qui ont des caractéristiques communes.

Pour conclure cette section, retenons que dans le Web sémantique les terminologies ont pour support soit des ontologies exprimées en OWL ou RDFS, soit des thésaurus exprimés en SKOS ou en une ontologie incluant une articulation avec des terminologies (internes ou externes), comme le CIDOC CRM ou Schema.org, ceci à l'exception de Wikidata qui suit une troisième voie de représentation. Ainsi, pour trouver les termes d'un graphe de connaissances, il faut examiner les ontologies et les thésaurus qui y sont utilisés. Nos expérimentations démontrent que repérer les ontologies et thésaurus doit même être la première étape de toute recherche de termes.

4. DÉTECTION DE RESSOURCES TERMINOLOGIQUES

Notre objectif est de trouver les termes, donc les ressources terminologiques utilisées dans les graphes de connaissances. C'est pourquoi dans cette section nous recherchons les thésaurus et les ontologies. Dans la section suivante, nous nous intéresserons plus particulièrement aux termes de thésaurus dans la mesure où il est très simple de trouver dans les graphes les classes et propriétés des ontologies utilisées, en utilisant des propriétés telles que `rdf:type`, `rdfs:subClassOf`, `owl:subClassOf`, ou `wdt:P31`, `wdt:P279`, ce que font déjà de nombreux outils [5, 11, 13, 23, 24].

4.1. PROTOCOLE EXPÉRIMENTAL

COMPOSITION DE LA LISTE DES GRAPHES ÉTUDIÉS. — La liste des graphes étudiés est disponible dans le dépôt Git lié à cet article ⁽²⁶⁾. Partant de la liste des graphes considérés dans IC 2022 [20], les graphes qui n'étaient plus accessibles ont été retirés, les adresses de points d'accès SPARQL ont été actualisées et une dizaine de graphes rendus disponibles entre temps a été ajoutée. Les graphes ne pouvant pas être interrogés automatiquement ont été retirés. Enfin, les graphes rendus inaccessibles au cours de l'expérimentation ont été retirés. Ainsi, les graphes restants sont ceux dont l'interrogation a été possible à chaque étape de l'expérimentation, soit entre janvier et mars 2024. Nous avons décidé de conserver deux graphes qui ne renvoient qu'une réponse partielle. Ainsi, 23 graphes ont été conservés sur les 37 graphes identifiés en premier lieu. Ils sont présentés dans le tableau 4.1.

⁽²⁶⁾<https://scm.univ-tours.fr/ovide/roia-ressources>

TABLE 4.1. Liste des graphes de connaissances utilisés.

Nom	Adresse endpoint	Sujet	Triplets	Voc.
Arkeogis - Hydromed (OpenArchaeo)	http://openarchaeo.huma-num.fr/federation/sparql	Données d'inventaire des aménagements hydrauliques romains en Italie du Nord (projet HYDROMED)	4 053	4
Arkeogis - Verduron (OpenArchaeo)	http://openarchaeo.huma-num.fr/federation/sparql	Données de fouille de Verduron et sites contemporains autour de Marseille grecque datant du III ^e s. av. J.-C.	523	4
Arsol (OpenArchaeo)	http://openarchaeo.huma-num.fr/federation/sparql	Données de fouilles archéologiques du Laboratoire Archéologie et Territoire de Tours	669 100	6
Chronique des fouilles en ligne (OpenArchaeo)	http://openarchaeo.huma-num.fr/federation/sparql	Données des opérations de terrain ayant eu lieu en Grèce et à Chypre	445 292	2
CoGhent	https://stad.gent/sparql	Données de la ville de Gand	198 894 439	4
Conservatoire 3D (OpenArchaeo)	http://openarchaeo.huma-num.fr/federation/sparql	Données de numérisation et/ou de restitution d'objets ou de sites archéologiques	12 216	2
Cultura Italia	http://dati.culturaitalia.it/sparql	Données de patrimoine culturel italien	41 901 551	4
DOREMUS	http://data.doremus.org/sparql	Données d'archives musicales françaises (BnF, Radio France, Philharmonie de Paris)	91 093 032	9
Epicherchell (OpenArchaeo)	http://openarchaeo.huma-num.fr/federation/sparql	Corpus des inscriptions antiques de Césarée de Maurétanie (Cherchell, Algérie)	3 251	2
Iceramm (OpenArchaeo)	http://openarchaeo.huma-num.fr/federation/sparql	Données sur lots céramiques typochronologiques	32 347	2
Kerameikos	http://kerameikos.org/sparql	Données de poterie	289 596	1
Kition-pervolia (OpenArchaeo)	http://openarchaeo.huma-num.fr/federation/sparql	Données de fouilles du site de Kition-Pervolia	32 287	3
Letter Sampo	http://ldf.fi/ckcc/sparql	Environ 200 000 lettres du Huygens Institute (projet LetterSampo)	505 942	2
LINCS	https://portal.lincsproject.ca/docs/tools/sparql/yasgui	Données du gouvernement canadien	5 941 533	11
Middle Kingdom	https://pnm.uni-mainz.de/sparql	Données sur le Moyen Empire Égyptien	1 539 752	1
Nomisma	http://nomisma.org/sparql	Données de numismatique	11 427 182	2
Outagr (OpenArchaeo)	http://openarchaeo.huma-num.fr/federation/sparql	Données sur l'outillage agricole Gallo-Romain	117 200	2
Portail archeoViz (OpenArchaeo)	http://openarchaeo.huma-num.fr/federation/sparql	Données sur les utilisations de l'application Archeoviz	3 216	1
Rita (OpenArchaeo)	http://openarchaeo.huma-num.fr/federation/sparql	Données des sources visuelles archéologiques de Tours	57 388	3
Smithsonian Museum	https://triplydb.com/smithsonian/american-art-museum/sparql/american-art-museum	Données sur le fonds du Smithsonian American Art Museum	2 802 768	2
Solidar (OpenArchaeo)	http://openarchaeo.huma-num.fr/federation/sparql	Données d'occupation humaine des massifs de Blois, Russy, Boulogne et Chambord des origines à nos jours à partir d'une campagne LiDAR (Light Detection And Ranging)	46 918	2
WarSampo	http://ldf.fi/warsa/sparql	Données sur la Seconde Guerre Mondiale en Finlande	14 385 118	4
WWI LOD	http://ldf.fi/wwi1od/sparql	Données sur la Première Guerre Mondiale	63 828	9

INDICES RECHERCHÉS. — Nos indices de base sont les déclarations explicites d'ontologie (`owl:Ontology`) ou de thésaurus (`skos:ConceptScheme`). La présence de telles déclarations repose sur le fait que les ressources terminologiques soient définies dans le schéma du graphe de connaissance analysé. Comme ce n'est pas toujours le cas, nous avons également procédé à une recherche manuelle des ressources terminologiques des graphes étudiés.

Pour les `skos:ConceptScheme` (resp. `owl:Ontology`), nous utilisons une requête SPARQL comptant le nombre de sujets distincts vérifiant le triplet `?s a skos:ConceptScheme` (resp. `?s a owl:Ontology`). Nous étiquetons manuellement chaque entité renvoyée selon s'il s'agit d'une ressource terminologique ou pas.

Afin d'établir une liste de référence des ressources terminologiques présentes dans ces graphes pour la comparer aux résultats obtenus grâce aux déclarations précédentes, nous avons également extrait toutes les URIs des entités des graphes. Nous en avons retiré la partie identifiante et nous les avons rassemblés pour établir la liste des domaines utilisés dans les URIs de ces graphes. L'extraction de l'intégralité des URIs des graphes étudiés est faite avec la requête suivante :

```
1 SELECT DISTINCT ?x
2 WHERE {
3   ?s ?p ?o .
4   BIND (REPLACE(STR(?s), "^(^.*)([/]*)*$", "$2") as ?x) .
5 }
```

Cette requête permet de récupérer les URI de toutes les entités en leur retirant la partie après le dernier slash. Nous automatisons la récupération des résultats de cette requête sur l'intégralité des graphes considérés. Ensuite nous récupérerons uniquement la partie de l'URI située avant le premier slash afin de retirer les doublons. Enfin, nous procédons à un étiquetage manuel de ces URI en les classifiant comme suit :

- URI : le début de l'URI correspond à un thésaurus qui pourra servir à l'indice Préfixe Répertoire présenté en section 5. Exemple : `https://ark.frantiq.fr/`
- le début de l'URI ne correspond pas à un thésaurus. Dans ce cas, nous avons les sous-types suivants :
 - ontologie : début d'URI d'ontologie. Exemple : `http://schema.org/`
 - référence : début d'URI d'entités d'un autre graphe, utilisé comme autorité dans le domaine. Exemple : `https://sws.geonames.org/`
 - lien : début d'URI de site Web. Exemple : `http://www.allocine.fr/`
 - exemple : début d'URI ne correspondant pas à un nom de domaine. Exemple : `http://foo.example/`

Cet étiquetage est fait par déréférencement, en analysant manuellement chaque entrée pour la classifier. Le résultat de ce travail est rendu public pour être réutilisable.

4.2. RÉSULTATS ET ANALYSE

4.2.1. Présentation des résultats

Avec la méthode des URIs, nous obtenons une liste de 521 ressources terminologiques potentielles, associées aux graphes dans lesquels elles ont été trouvées. Parmi elles nous en identifions 19 comme étant des thésaurus, dont les noms de domaine sont listés en table 4.2. Ce sont :

- Des thésaurus spécifiques aux graphes dans lesquels ils sont utilisés, par exemple <http://culturaitalia.it/> pour Cultura Italia. Ces noms de domaine ne sont pas réutilisés dans d'autres graphes et n'ont probablement pas vocation à l'être.
- Des ressources terminologiques connues et utilisées dans des domaines spécifiques et dont l'utilisation relève souvent de la bonne pratique, par exemple <http://www.mimo-db.eu/> pour DOREMUS. Ces ressources sont alors destinées à être partagées et réutilisées.
- trois ressources terminologiques axées sur les métadonnées, par exemple <http://IFLStandards.info/>. Ces ressources terminologiques ne nous renseignent alors pas tant sur le contenu du graphe mais sur la façon dont les données sont encodées, partagées et, par la suite, mobilisables.

TABLE 4.2. Liste des thésaurus détectés dans chaque graphe de connaissances.

Graphe de connaissances	Ressource terminologique (noms de domaines)
Arkeogis – Hydromed	https://ark.frantiq.fr/
Arkeogis – Verduron	https://ark.frantiq.fr/
Arsol	https://ark.frantiq.fr/ ; http://vocab.Getty.edu/
Chronique des fouilles en ligne	https://ark.frantiq.fr/ ; http://vocab.Getty.edu/
GoGhent	https://stad.gent/ ; http://stad.gent/ (S); https://data.vlaanderen.be/ (S)
Conservatoire 3D	https://ark.frantiq.fr/ ; http://vocab.Getty.edu/
Cultura Italia	http://culturaitalia.it/ (S)
DOREMUS	http://www.mimo-db.eu/ (S); http://data.doremus.org/ (S, O); http://rdaregistry.info/ (S); http://IFLStandards.info/ ; http://metadataregistry.org/
Epicherchell	https://ark.frantiq.fr/ ; http://vocab.Getty.edu/
Iceramm	https://ark.frantiq.fr/ ; http://vocab.Getty.edu/
Kerameikos	http://kerameikos.org/ (S, O)
Kition pervolia	https://ark.frantiq.fr/ ; http://vocab.Getty.edu/ ; https://ark.mom.fr/
LINCS	http://vocab.Getty.edu/ ; http://id.lincsproject.ca/ (S); http://IFLStandards.info/ ; https://vocab.Getty.edu/
Middle Kingdom	https://pnm.uni-mainz.de/ (S)
Nomisma	http://nomisma.org/ (S, O)
Outagr	https://ark.frantiq.fr/ ; http://vocab.Getty.edu/
Portail ArcheoViz	https://ark.frantiq.fr/
Rita	https://ark.frantiq.fr/ ; http://vocab.Getty.edu/ ; https://ark.mom.fr/
Smithsonian Museum	http://data.americanart.si.edu/ (O)
Solidar	https://ark.frantiq.fr/ ; http://vocab.Getty.edu/
WarSampo	http://ldf.fi/ (S, O)
WWILOD	http://bibframe.org/vocab/ (O); http://ldf.fi/

La table 4.2 présente la liste des graphes de connaissances en leur associant les thésaurus détectés par la méthode des URIs. Les mentions S et O indiquent que les thésaurus sont détectés respectivement via la méthode `skos:ConceptScheme` et `owl:Ontology`.

Nous remarquons que l'utilisation des indices `skos:ConceptScheme` et `owl:Ontology` est souvent liée, bien que ces deux indices ne renvoient pas les mêmes réponses. Quand `skos:ConceptScheme` et `owl:Ontology` sont utilisés en même temps, les résultats des requêtes leur correspondant ne se recoupent que partiellement ; les domaines renvoyés par la requête `skos:ConceptScheme` sont alors contenus dans ceux renvoyés par la requête `owl:Ontology`. Lorsqu'elle renvoie une réponse, la requête `skos:ConceptScheme` renvoie systématiquement un thésaurus spécifique au graphe étudié. La requête `owl:Ontology` renvoie des ontologies utilisées pour constituer la T-box du graphe étudié. Ce sont les mêmes ontologies qui reviennent : OWL (pour 5 des 11 graphes ne renvoyant pas une réponse vide), SKOS (3 réponses), CRM (4 réponses), etc. On remarque également la présence d'ontologies plus spécifiques à certains graphes uniquement : PROV, DCAT, FOAF, etc. Nous remarquons que ces requêtes renvoient pour les graphes Virtuoso des noeuds vides qui correspondent probablement à une structure spécifique. Enfin elles renvoient des réponses vides pour les graphes OpenArcheo, car ils ne contiennent pas leur schéma (TBox).

Environ trois quart des ressources terminologiques identifiées sont spécifiques aux graphes dans lesquels elles apparaissent. Représenter dans la table 4.2 les ressources trouvées par leur nom de domaine ne permet pas de montrer deux ressources terminologiques différentes ayant le même domaine (par exemple, `kerameikos.org/id` et `kerameikos.org/editor` sont alors confondus). De la même manière, cela ne nous permet pas de distinguer deux graphes différents qui utiliseraient le même nom de domaine pour renvoyer à des ressources terminologiques différentes. C'est le cas de WarSampo et WWILOD, qui renvoient respectivement à `http://ldf.fi/warsa` et à `http://ldf.fi/schema/warsa/` pour l'un et à `http://ldf.fi/ww1lod/` pour l'autre.

COMPARAISON DES MÉTHODES. — En recoupant les méthodes, nous obtenons une liste de 19 vocabulaires. La méthode par URIs les extrait tous, contrairement à la méthode de recherche de déclaration explicite. Lorsque la même ressource terminologique est détectée par plusieurs méthodes dans la table 4.2, il s'agit de cas où le nom de domaine sert à la fois à un thésaurus et à une ontologie (par exemple, pour WarSampo `http://ldf.fi/warsa` et `http://ldf.fi/schema/warsa/` sont confondus).

4.2.2. *Analyse des ressources terminologiques trouvées*

QUANTIFICATION DE L'UTILISATION DES RESSOURCES TERMINOLOGIQUES IDENTIFIÉES. Les graphes d'OpenArcheo reposant sur les mêmes spécifications, ils utilisent logiquement les mêmes vocabulaires. Nous remarquons la double utilisation de PAC-TOLS et PACTOLS2 pour Kition-pervolia et Rita, qui correspond probablement à un défaut d'alignement. 10 des 23 graphes utilisent leurs propres vocabulaires (3 vocabulaires pour CoGhent). Quatre vocabulaires sont utilisés sur plusieurs graphes :

PACTOLS/PACTOLS2, le Getty et l'IFLA. Enfin, LINC3 nous semble particulièrement intéressant dans le fait qu'il fait référence à deux vocabulaires utilisés dans d'autres graphes en plus d'utiliser le sien : le Getty, utilisé dans les graphes d'OpenArcheo, et IFLA, utilisé dans DOREMUS. Dans les graphes OpenArcheo, le Getty semble être utilisé en déclaration de correspondance à des éléments du PACTOLS. Nous remarquons cependant que cette équivalence est déjà décrite dans les termes du PACTOLS, ce qui pourrait indiquer une duplication d'informations. Dans LINC3, le Getty est utilisé comme vocabulaire à part entière.

QUELS SONT LES TERMES UTILISÉS AU SEIN DE CES RESSOURCES TERMINOLOGIQUES ?
Nous nous concentrons ici sur les termes eux-mêmes utilisés au sein des ressources terminologiques identifiées comme partagées au sein de différents graphes lors de l'expérimentation. Deux cas de figure existent : celui des graphes OpenArcheo qui partagent les mêmes standards de description et celui de LINC3 et DOREMUS, qui utilisent leur propre ressource terminologique spécifique tout en faisant également usage de ressources terminologiques partagées, notamment la référence à IFLA Standards.

PACTOLS. — Le vocabulaire contrôlé PACTOLS est issu de l'archéologie. Il est utilisé dans l'intégralité des graphes OpenArcheo. Étant publiés par la même organisation, ces graphes obéissent aux mêmes standards, et sont donc considérés comme comparables de ce point de vue. Nous avons comparé les 10 premiers termes de ce vocabulaire utilisés dans chacun de ces graphes. Ces 10 premiers termes ne se recoupent que très peu entre les graphes. Cela montre que les termes sont très spécifiques aux graphes dans lesquels ils sont trouvés, et très spécifiques au sujet de ces graphes. Les graphes d'OpenArcheo partagent tous le même schéma, mais portent tous sur des sujets différents au sein du domaine archéologique, il est normal d'y retrouver des termes très différents.

IFLA Standards. — La ressource terminologique IFLAstandards représente des méthodes de description bibliographique. Les utilisations faites de cette ressource terminologique dans LINC3 et dans DOREMUS diffèrent : le premier utilise la partie FRBRoo d'IFLAstandards alors que le second utilise l'UNIMARC.

Nous pouvons tirer deux conclusions à partir de ces observations dans l'usage de PACTOLS et de IFLA Standards. Premièrement, des graphes de connaissances partageant le même vocabulaire n'en utilisent pas les mêmes parties. Deuxièmement, les termes caractérisent finement le contenu des graphes de connaissances. Par exemple, pour les différents graphes d'OpenArcheo, les termes sont particulièrement importants pour qualifier le contenu de chacun des graphes, d'autant que l'ontologie utilisée (le CIDOC CRM) est volontairement générique.

5. DÉTECTION DE TERMES

5.1. PROTOCOLE DES EXPÉRIMENTATIONS

Nous avons utilisé trois indices pour la détection des termes, résumés dans la table 5.1, 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).

TABLE 5.1. Trois indices pour la détection des termes.

Indice	Description	Précision	Rappel	Prérequis	Requête
propriété répertoriée	Sujet d'une propriété en principe dédiée à la terminologie comme <code>skos:prefLabel</code> , etc.	faible	variable	élevé	Q1
classe répertoriée	Instance d'une classe dédiée à la terminologie comme <code>skos:Concept</code> , <code>crm:E55_Type</code> , etc.	très élevée	variable	élevé	Q2
Préfixe répertorié	Préfixe d'URI correspondant à un thésaurus répertorié, comme PACTOLS ou AAT du Getty	très élevée	élevé	très élevé	Q3

PROPRIÉTÉ RÉPERTORIÉE. — Les sujets de certaines propriétés ont une forte chance d'être des concepts dénotant des termes. Nous avons d'abord pensé que ce serait 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 encore les propriétés P127 `has broader term` et P150 `defines typical parts of` pour le CIDOC CRM. Mais, comme pour les classes répertoriées, ces propriétés peuvent être absentes du graphe analysé et être plutôt dans le graphe où est définie la terminologie. La **requête Q1** utilisée dans les expérimentations pour cet article est la suivante :

```

1  SELECT DISTINCT ?s WHERE {
2    ?s skos:prefLabel ?o.
3  }
```

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, `ontolex:LexicalConcept` pour OntoLex-lemon et Termlex ou `schema:DefinedTerm` pour Schema.org. 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 comprendre l'ontologie de référence du domaine (par exemple CIDOC CRM pour savoir que la classe E55 Type a pour instances des concepts de terminologies). La **requête Q2** utilisée dans les expérimentations pour cet article est :

```

1  SELECT DISTINCT ?s {
2    {
3      SELECT ?s WHERE {
4        ?s a crm:E55_Type
5      }
6    }
7    UNION
8    {
9      SELECT ?s WHERE {
10        ?s a skos:Concept
11      }
12    }
13  }
14
```

PRÉFIXE RÉPERTORIÉ. — Une méthode simple est de rechercher les URIs correspondant à des thésaurus connus. Par exemple, toutes les URIs débutant par <https://ark.frantiq.fr/> appartiennent au thésaurus PACTOLS⁽²⁷⁾. La précision de cette approche est très élevée, mais sa couverture dépend du fait que le graphe de connaissances ciblé utilise des thésaurus connus ou pas. Cette approche ne permet pas la découverte de nouveaux thésaurus, ce qui la rend peu adaptée aux graphes de connaissances nouveaux. C'est donc ici que les expérimentations de la section 4 prennent tout leur intérêt : les préfixes utilisés dans les expérimentations qui suivent sont ceux identifiés en amont lors de l'étape de détection des ressources terminologiques. La **requête Q3** utilisée dans les expérimentations pour cet article est la suivante, où x correspond à l'un des préfixes connus :

```

1  SELECT DISTINCT ?s WHERE {
2    ?s ?p ?o .
3    FILTER (STRSTARTS(STR(?s), x))
4  }
```

Cette requête légèrement modifiée permet aussi de détecter des préfixes connus dans les objets des triplets.

À noter que dans l'article d'origine [20], une requête Universaux Filtrés avait été présentée. Elle visait à retrouver les termes en s'appuyant sur des indices topologiques dans le graphe. La campagne d'expérimentations menée pour cet article l'ayant évaluée comme peu féconde, elle a été retirée de nos propositions. Elle ajoutait du bruit dans les résultats sans détecter de termes qu'aucune des trois autres méthodes n'aurait retourné.

Les 23 graphes ont été interrogés avec un script Python utilisant la librairie SPARQLWrapper⁽²⁸⁾. Le code utilisé est disponible dans le dépôt Git <https://scm.univ-tours.fr/ovide/roia-ressources> lié à cet article.

⁽²⁷⁾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.

⁽²⁸⁾<https://sparqlwrapper.readthedocs.io/en/latest/>

Dans un deuxième temps, pour pousser nos analyses, nous avons choisi 3 graphes, Kerameikos, Cultura Italia et Epicherchell. Les deux premiers sont choisis pour des questions de taille et de disponibilité. Le dernier l'est parce qu'il appartient à l'ensemble des graphes d'OpenArcheo, qui représente 12 des 23 graphes que nous étudions. Pour chacun, les réponses des trois requêtes sont étudiées et comparées. Notre but est de voir si certaines méthodes sont plus efficaces que d'autres dans la détection de termes, et de déterminer à quel point chaque requête recoupe ou non les autres. Pour cela, nous créons une liste de toutes les entités renvoyées par les trois requêtes de détection des termes et nous indiquons pour chacune d'entre elles par quelle requête elle a été repérée.

Enfin, nous avons étiqueté à la main ces entités selon leur statut de terme de thésaurus ou non, en suivant les définitions amenées précédemment. Lorsqu'il n'est pas un terme, nous catégorisons ce faux positif de la manière suivante :

- référence externe : l'entité appartient à un autre graphe.
 - type autorité : l'entité est une entité nommée de référence, par exemple des renvois à Wikidata ou à GeoNames
- référence interne : l'entité appartient au graphe examiné
 - type général : l'entité fait référence à une autre entité du graphe, généralement à un pointeur
 - type ressource terminologique : l'entité fait référence à un élément d'ontologie ou de thésaurus du graphe, mais sans être un terme, par exemple <https://w3id.org/whow/controlled-vocabulary/chemical-substances> qui est le nom d'un vocabulaire, pas d'un terme de ce vocabulaire
- erreurs de constitution du graphe : l'entité est un élément qui ne devrait pas apparaître dans le graphe, par exemple des URIs localhost

5.2. RÉSULTATS ET ANALYSE

5.2.1. *Quantification des résultats des requêtes pour l'ensemble des graphes*

Le tableau 5.2 présente le nombre de termes découverts par chacune des méthodes du tableau 5.1.

On remarque qu'un usage immodéré de la propriété `skos:prefLabel` rend illusoire l'utilisation de la méthode « propriété répertoriée » pour les graphes de OpenArcheo. Dans plusieurs graphes de OpenArcheo, cette méthode ne donne rien parce que les concepteurs ont associé aux concepts du thésaurus PACTOLS un autre terme que dans le thésaurus 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 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 du jeu de données original.

Ce qui ressort de façon évidente de ces résultats, c'est que ces termes représentent bien une partie quantitativement importante des graphes analysés.

TABLE 5.2. Résultats des requêtes de détection de termes.

Nom du graphe	propriété répertoriée	classe répertoriée	Préfixe répertorié
Arkeogis - Hydromed	464	0	100
Arkeogis - Verduron	45	0	30
Arsol	52 417	0	189
Chronique des fouilles en ligne	55 840	0	164
CoGhent	135 223	6 352 411	5 320 715
Conservatoire 3D	1 570	0	224
Cultura Italia	875	866	865
DOREMUS	12 213	12 212	3 004 466 (réponse partielle)
Epicherchell	321	0	50
Iceramm	4 409	0	43
Kerameikos	461	460	1 181
Kition pervolia	2 144	0	1 634
Letter Sampo	61 787	0	0
LINCS	4 039	4 333	time out
Middle Kingdom	28 205	3 595	2443 63
Nomisma	164 186	150 975	44 779
Outagr	18 072	0	698
Portail archeoViz	405	0	68
Rita	7 440	0	1 027
Smithsonian Museum	16	0	40 879
Solidar	5 128	0	94
WarSampo	1 769 137	7 090	1 811 167
WWILOD	9 203	0	8 817

5.2.2. Qualification des résultats sur trois graphes en particulier

Dans la table 5.3 la requête Q1 correspond à la requête de propriété répertoriée, la requête Q2 à celle de classe répertoriée et Q3 à celle de préfixe répertorié.

TABLE 5.3. Résultats par requête et par graphe et catégories des faux positifs.

label	catégorie	Kerameikos			Cultura Italia			Epicherchell			Whow		
Faux positif		Q1	Q2	Q3	Q1	Q2	Q3	Q1	Q2	Q3	Q1	Q2	Q3
	Erreurs	0	0	0	9	0	0	0	0	0	0	0	0
	Ref. externe : autorité	0	0	0	0	0	0	1	0	0	0	0	0
	Ref. interne : générale	0	0	705	0	0	0	320	0	0	0	0	0
	Ref. interne : ressource terminologique	0	0	2	1	1	0	0	0	0	0	0	3
Total des faux positifs		0	0	707	10	1	0	321	0	0	0	0	3
Terme	Terme	461	460	447	865	865	865	0	0	50	1096	1096	1096
Total termes identifiés		461	460	447	865	865	865	0	0	50	1096	1096	1096
Total général		461	460	1154	875	866	865	321	0	50	1096	1096	1099
Total termes du graphe		464	464	464	865	865	865	50	50	50	1096	1096	1096

KERAMEIKOS. — Les requêtes sur le graphe Kerameikos renvoient 707 faux positifs en tout pour 464 termes. Ces faux positifs sont principalement des références internes générales où le graphe fait référence à lui-même, par exemple avec des réponses comme <http://kerameikos.org/id/bowl#provenance> (705 entités sur 707 faux positifs). À la marge, on comptabilise également deux références internes de ressource terminologique comme <http://kerameikos.org/id/>, qui renvoie certes à la ressource terminologique du graphe mais qui n'est pas un terme. Les requêtes de propriété répertoriée et de classe répertoriée sont particulièrement efficaces sur ce graphe, puisqu'elles se recoupent quasi parfaitement (la requête de propriété répertoriée renvoie une entité de plus que celle de classe répertoriée) et renvoient à elles deux 460 termes sur les 464 termes du graphe. Les 4 termes manquants correspondent à des termes recouverts par l'ontologie interne de Kerameikos ; ils sont renvoyés par la requête de préfixe répertorié. L'intégralité des 707 faux positifs renvoyés sont identifiés par la requête de préfixe répertorié, polluée par les références internes avec pointeurs. La requête optimale pour identifier les termes sur ce graphe semble être la requête de propriété répertoriée. La requête de préfixe répertorié permet cependant d'ajouter 4 termes qui n'avaient pas été identifiés par cette première requête, bien qu'elle apporte beaucoup de faux positifs.

CULTURA ITALIA. — Les requêtes sur le graphe Cultura Italia renvoient 10 faux positifs pour 865 termes. Ces faux positifs sont principalement de la catégorie *autre erreur de constitution des graphes*. Ils prennent la forme de noeuds vides, sauf le dernier qui est une référence interne de ressource terminologique dont l'URI est inaccessible : <http://dati.culturaitalia.it/risorse1>. Il peut s'agir d'un terme mal défini ou déprécié même si cela est peu probable au vu du domaine dati.culturaitalia.it. La catégorisation en faux positif se justifie par l'inaccessibilité de l'URI, associée à sa non lisibilité ("ressorse1"). Les requêtes de propriété répertoriée, de classe répertoriée et de préfixe répertorié renvoient toutes 865 termes pour respectivement 10, 1 et 0 faux positifs. 10 faux positifs sont ramenés par la requête de propriété répertoriée, parmi lesquels 9 erreurs de constitution du graphe et la référence interne de ressource terminologique décrite plus haut. La requête de classe répertoriée renvoie la même référence interne erronée que celle de la requête de propriété répertoriée.

EPICHERCHELL. — Les requêtes sur le graphe Epicherchell renvoient 321 faux positifs pour 50 termes identifiés. Ces faux positifs sont essentiellement des références internes à des pointeurs ou des objets particuliers (320 résultats) comme <http://ccj-epicherchell.huma-num.fr/interface/fiche.php?id=21>. Le dernier faux positif (requête de propriété répertoriée) est une référence externe à GeoNames : <https://sws.geonames.org/2501440/>. La requête la plus efficace pour ce graphe est la requête de préfixe répertorié, puisqu'elle renvoie l'intégralité des termes pour aucun faux positif. L'intégralité des faux positifs sont ramenés par la requête de propriété répertoriée. L'absence de réponse correcte des requêtes de propriété répertoriée et de classe répertoriée s'explique par le fait que le schéma du graphe Epicherchell n'est pas accessible via le point d'accès public.

5.3. PROPOSITION D'UNE MÉTHODOLOGIE

En s'appuyant sur nos expérimentations, nous estimons que la meilleure méthodologie pour retrouver les termes d'un graphe est la suivante :

- (1) Identifier les ressources terminologiques du graphe en utilisant les indices `skos:ConceptScheme`, `owl:Ontology` et la recherche de noms de domaine si les premiers ne retournent rien, en testant déjà les 19 vocabulaires découverts dans ce travail.
- (2) À partir de ces ressources terminologiques, retrouver les termes en extrayant toutes les URI du graphe qui leur correspondent, par l'un des trois indices de la table 5.1.

Pour montrer la généralité de cette méthodologie au-delà du domaine du patrimoine culturel, nous avons expérimenté sur un graphe de connaissances d'un autre domaine, le graphe WHOW-KG [4] contenant des informations sur la qualité et la consommation de l'eau au niveau européen. Nous avons choisi ce graphe en raison de nos travaux pour l'ARD JUNON, dont le cœur de recherche porte sur la constitution de jumeaux numériques eau, air et sol pour répondre à des problématiques environnementales. Le point d'accès requêté est le suivant : <https://semscout.istc.cnr.it/sparql/>. Une fois les thésaurus identifiés en utilisant l'indice `skos:ConceptScheme`, nous avons lancé les trois requêtes de détection des termes pour comparer leur pertinence.

Les trois requêtes renvoient 1 096 termes pour 3 faux positifs. Ces faux positifs sont des références internes de ressource terminologique (3 résultats) comme <https://w3id.org/whow/controlled-vocabulary/chemical-substances>, qui correspond bien à une ressource terminologique mais qui n'est pas un terme. Les requêtes de propriété répertoriée et de classe répertoriée renvoient l'intégralité des termes du graphe, sans aucun faux positif. La requête de préfixe répertorié renvoie également les 1 096 termes avec 3 faux positifs, qui sont en réalité les trois noms des thésaurus utilisés dans ce graphe. Ainsi, notre méthodologie nous permet d'identifier l'intégralité des termes du graphe. Nous remarquons que, comme pour Kerameikos, les requêtes de propriété et de classe répertoriée se révèlent meilleures que la requête de préfixe répertorié.

Lorsque les ressources terminologiques sont incluses dans les graphes, les déclarations `skos:ConceptScheme` ou `owl:Ontology` permettent de les trouver très simplement. La méthode des URIs demandant une analyse manuelle qui peut être laborieuse, nous apportons aux concepteurs de graphes de connaissances une nouvelle recommandation : celle d'indiquer dans les métadonnées de leurs graphes les ressources terminologiques utilisées, conformément aux principes FAIR [27]. C'est possible en utilisant des propriétés spécifiques qui recensent les ressources terminologiques utilisées au sein d'un graphe. Cela inclut, sans s'y limiter, les propriétés `void:vocabulary`, `dct:conformsTo` et `dc:themeTaxonomy`.

6. 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 thésaurus, définis soit avec SKOS, soit avec d'autres vocabulaires. Dans le premier cas (ontologies) elles apparaissent dans les profils ou résumés de graphes de connaissances classiques, qui montrent les classes et les propriétés utilisées. Dans le deuxième cas (thésaurus), nous montrons qu'il est compliqué de les détecter. Nous avons testé plusieurs méthodes pour identifier des éléments de thésaurus dans 23 graphes de connaissances offrant des points d'accès SPARQL dans le domaine du patrimoine culturel. De ces expériences nous avons tiré une méthodologie, que nous avons validée sur un graphe d'un domaine différent, celui des sciences de la Terre. Ces travaux sont facilement reproductibles, notamment grâce à la liste étiquetée manuellement des URIs de termes détectés dans ces 23 graphes. L'automatisation complète de cette méthodologie pour son application à un nouveau corpus de graphes de connaissances est envisagée. À notre connaissance, les travaux qui annoncent recenser les « vocabulaires » utilisés dans les graphes de connaissances [11, 23, 24] en recensent uniquement les ontologies. Pourtant les thésaurus apportent autant d'interopérabilité et de sémantique, d'où l'intérêt d'une méthodologie pour en retrouver les termes dans les graphes. C'est déjà pour nous une base pour générer automatiquement des profils de graphes plus informatifs, cela pourrait également être exploité pour réécrire ou optimiser des requêtes, tout comme les informations sur les classes et les propriétés instanciées dans le graphe.

7. REMERCIEMENTS

Le travail présenté à IC 2022 a été financé par l'ANR-18-CE38-0009 SESAMES et sa présente extension par l'ARD Région Centre Val de Loire JUNON.

BIBLIOGRAPHIE

- [1] F. AMARGER, C. ROUSSEY, J.-P. CHANET, O. HAEMMERLÉ & N. HERNANDEZ, « État de l'art : Extraction d'information à partir de thésaurus pour générer une ontologie », in *INFORSID*, 2013, p. 29-44.
- [2] C. BEKIARI, G. BRUSEKER, M. DOERR, C.-E. ORE, S. STEAD & A. VELIOS, « Definition of the CIDOC Conceptual Reference Model. Last official version : 7.1.1 », Tech. report, ICOM/CIDOC Documentation Standards Group. CIDOC CRM SIG, 2021.
- [3] A. BRINKMANN, A. PRIMPEL & C. BIZER, « The Web Data Commons Schema.org Data Set Series », in *Companion Proceedings of the ACM Web Conference 2023*, WWW '23 Companion, Association for Computing Machinery, 2023, p. 136-139.
- [4] G. CARLETTI, E. GIULIANELLI, A. S. LIPPOLIS, G. LODI, A. G. NUZZOLESE, M. PICONE & G. SETTANTA, « The Water Health Open Knowledge Graph », <https://arxiv.org/abs/2305.11051>, 2023.
- [5] S. CEBIRIC, F. GOASDOUE, H. KONDYLAKIS, D. KOTZINOS, I. MANOLESCU, G. TROULLINOÛ & M. ZNEIKA, « Summarizing Semantic Graphs: A Survey », *The VLDB Journal* **28** (2018), p. 295-327.
- [6] M.-H. DANG, A. GAIGNARD, H. SKAF-MOLLI & M. PASCAL, « Schema.org: How is it used? », in *ISWC 2023, CEUR Workshop Proceedings*, CEUR Workshop Proceedings, vol. 3632, CEUR-WS.org, 2023.

- [7] L. DIOP, A. GIACOMETTI, B. MARKHOFF & A. 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)*, CEUR Workshop Proceedings, vol. 2949, CEUR-WS.org, 2021.
- [8] L. DIOP, B. MARKHOFF & A. SOULET, « TTProfiler: Types and Terms Profile Building for Online Cultural Heritage Knowledge Graphs », *Journal of Computing and Cultural Heritage* **16** (2023), n° 3, p. 1-22.
- [9] B. FARIAS LÓSCIO, C. BURLE & N. 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.
- [10] FRANTIQU, « Le Thésaurus PACTOLS », <https://www.frantiq.fr/pactols/le-thesaurus/>, 2020, Accessed on 2022-28-02.
- [11] J. FREY & S. HELLMAN, « FAIR Linked Data – Towards a Linked Data Backbone for Users and Machines », in *Companion Proceedings of the Web Conference 2021*, WWW '21, Association for Computing Machinery, 2021, p. 431–435.
- [12] F. GANDON, C. FARON-ZUCKER & O. CORBY, *Le Web sémantique*, Dunod, 2012.
- [13] F. GOASDOUÉ, P. GUZEWICZ & I. MANOLESCU, « RDF graph summarization for first-sight structure discovery », *VLDB J.* **29** (2020), n° 5, p. 1191-1218.
- [14] R. V. GUHA, D. BRICKLEY & S. MACBETH, « Schema.org: evolution of structured data on the web », *Communications of the ACM* **59** (2016), n° 2, p. 44-51.
- [15] A. HOGAN, E. BLOMQUIST, M. COCHEZ, C. D'AMATO, G. DE MELO, C. GUTIÉRREZ, S. KIRrane, J. E. LABRA GAYO, R. NAVIGLI, S. NEUMAIER, A. NGONGA NGOMO, A. POLLERES, S. M. RASHID, A. RULA, L. SCHMELZEISEN, J. F. SEQUEDA, S. STAAB & A. ZIMMERMANN, « Knowledge Graphs », *ACM Comput. Surv.* **54** (2021), n° 4, article no. 71 (37 pages).
- [16] D. KLESS, L. JANSEN & S. MILTON, « A content-focused method for re-engineering thesauri into semantically adequate ontologies using OWL », *Semantic Web* **7** (2016), n° 5, p. 543-576.
- [17] D. KLESS, S. MILTON, E. KAZMIERCZAK & J. LINDENTHAL, « Thesaurus and ontology structure : Formal and pragmatic differences and similarities », *Journal of the Association for information science and technology* **66** (2015), n° 7, p. 1348-1366.
- [18] M. LAMÉ, P. PITTET, F. PONCHIO, B. MARKHOFF & E. 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, ODOCHCAiSE 2019*, CEUR Workshop Proceedings, vol. 2375, CEUR-WS.org, 2019, p. 37-48.
- [19] B. LANGLET, « Universaux (GP), dans Maxime Kristanek (Dir.), l'Encyclopédie philosophique », <https://encyclo-phil.fr/universaux-gp>, 2019, Consulté le 15/03/2022.
- [20] B. MARKHOFF & A. SOULET, « Où sont les termes ? », in *Ingénierie des Connaissances IC 2022*, 2022.
- [21] E. PERRIN, « Thésaurus et interopérabilité des données archéologiques : le projet HyperThesau », *Humanités numériques* **4** (2021).
- [22] C. ROCHE & M. PAPADOPOULOU, « Rencontre entre une philologue et un terminologue au pays des ontologies », *Revue Ouverte d'Intelligence Artificielle* **1** (2020), n° 1, p. 43-70.
- [23] Q. SHI, J. WANG, J. Z. PAN & G. CHENG, « VOYAGE: A Large Collection of Vocabulary Usage in Open RDF Datasets », in *The Semantic Web – ISWC 2023*, Springer Nature Switzerland, 2023, p. 211-229.
- [24] B. SPAHIU, R. PORRINI, M. PALMONARI, A. RULA & A. MAURINO, « ABSTAT: Ontology-Driven Linked Data Summaries with Pattern Minimalization », in *The Semantic Web – ESWC 2016 Satellite Events, Revised Selected Papers*, Springer, 2016, p. 381-395.
- [25] F. SUCHANEK, M. ALAM, T. BONALD, P.-H. PARIS & J. SORIA, « Integrating the Wikidata Taxonomy into YAGO », <https://arxiv.org/abs/2308.11884>, 2023.
- [26] D. VRANDEČIĆ, L. PINTSCHER & M. KRÖTZSCH, « Wikidata: The Making Of », in *Companion Proceedings of the ACM Web Conference 2023*, WWW '23 Companion, Association for Computing Machinery, 2023, p. 615-624.
- [27] M. D. WILKINSON, M. DUMONTIER, I. J. AALBERSBERG, G. APPLETON, M. AXTON, A. BAAK, N. BLOMBERG, J.-W. BOITEN, L. B. DA SILVA SANTOS, P. E. BOURNE, J. BOUWMAN, A. J. BROOKES, T. CLARK,

M. CROSAS, I. DILLO, O. DUMON, S. EDMUNDS, C. T. EVELO, R. FINKERS, A. GONZALEZ-BELTRAN, A. J. G. GRAY, P. GROTH, C. GOBLE, J. S. GRETHE, J. HERINGA, P. A. C. 'T HOEN, R. HOOFT, T. KUHN, R. KOK, J. KOK, S. J. LUSHER, M. E. MARTONE, A. MONS, A. L. PACKER, B. PERSSON, P. ROCCA-SERRA, M. ROOS, R. VAN SCHAİK, S.-A. SANSONE, E. SCHULTES, T. SENGSTAG, T. SLATER, G. STRAWN, M. A. SWERTZ, M. THOMPSON, J. VAN DER LEI, E. VAN MULLIGEN, J. VELTEROP, A. WAAGMEESTER, P. WITTENBURG, K. WOLSTENCROFT, J. ZHAO & B. MONS, « The FAIR Guiding Principles for scientific data management and stewardship », *Scientific Data* **3** (2016), n° 1, article no. 160018.

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, i.e. words associated with concepts defined in thesauri, can also inform about its content, as much as the class names. Based on 23 SPARQL Endpoints, we present different techniques for automatically retrieving terms that are not class names. The results obtained by each technique enable us to propose a methodology for finding terms present in a knowledge graph. Experimental results and queries supporting the methodology are available online.

KEYWORDS. — CIDOC CRM, knowledge graph, ontology, SPARQL, term, terminology, thesaurus.
