Ontologie des marques de domaines appliquée aux dictionnaires de langue générale[1]

Rute COSTA
NOVA CLUNL, Centro de Linguística da Universidade NOVA de Lisboa
Sara CARVALHO
NOVA CLUNL, Centro de Linguística da Universidade NOVA de Lisboa, CLLC, Centro de Línguas, Literaturas e Culturas da Universidade de Aveiro
Ana SALGADO
NOVA CLUNL, Centro de Linguística da Universidade NOVA de Lisboa, Academia das Ciências de Lisboa, Instituto de Lexicologia e Lexicografia da Língua Portuguesa, R. da Academia das Ciências de Lisboa
Alberto SIMÕES
2Ai – School of Technology, IPCA
Toma TASOVAC
Belgrade Center for Digital Humanities

Résumé

Dans cet article, nous présentons OntoDomLab-Med, une ontologie des marques de domaines des sciences médicales et de la santé. Nous avons élaboré une taxonomie à partir des marques présentes dans la liste des abréviations du *Dicionário da Língua Portuguesa Contemporânea* de l'Académie des Sciences de Lisbonne. Notre objectif est de mettre en rapport OntoDomLab-Med et les entrées sélectionnées du dictionnaire balisées en TEI Lex-0 – système de balisage plus stricte et plus adapté que TEI au codage des dictionnaires – en ligne avec les principes FAIR. L'ontologie construite avec Protégé et codifiée en OWL permet l'exportation des connaissances dans un format d'échange interopérable permettant que l'ontologie puisse être appliquée à différentes ressources lexicales pour référencer les domaines indépendamment de la langue utilisée.

OntoDomLab-Med sera utile non seulement pour rechercher de l'information par domaine, mais permettra au lexicographe d'être plus cohérent dans son travail de chercheur et de codeur de l'information à des fins lexicographiques.

Mots-clés : lexicographie, ontologie de domaine, marques de domaines, annotation linguistique, standards.

[1] Recherche financée par la Fondation nationale portugaise à travers la FCT – Fondation pour la Science et Technologie - Centro de Linguística da Universidade NOVA de Lisboa – UID/LIN/03213/2020, par le Programme de recherche et d'innovation Horizon 2020 de l'Union européenne dans le cadre d'une convention de subvention N° 731015 (ELEXIS), et à travers FCT/MCTES comme partie du projet 2Ai – School of Technology, IPCA – UIDB/05549/2020.

Abstract

In this article, we present OntoDomLab-Med, a domain label ontology focused on medical and health sciences. We have developed a taxonomy from the labels included in the list of abbreviations of the *Dicionário da Língua Portuguesa Contemporânea* of the Lisbon Academy of Sciences. Our goal is to connect OntoDomLab-Med to a set of selected entries from the dictionary which have been encoded in TEI Lex-0, a stricter and more suitable format than TEI for dictionary encoding, in line with the FAIR principles. The ontology, built with Protégé and represented in OWL, allows the knowledge to be exported using an interoperable exchange format, thereby enabling the ontology to be applied to different lexical resources and to various domains, regardless of the natural language being used.

OntoDomLab-Med will be useful not only to research information by domain but it will also allow the lexicographer to be more consistent in his/her work as both researcher and coder of information for lexicographic purposes.

Keywords: lexicography, domain ontology, domain marks, linguistic annotation, standards.

Resumen

En este artículo, presentamos OntoDomLab-Med, una ontología de marcas de dominios de las ciencias médicas y de la salud. Hemos desarrollado una taxonomía a partir de las marcas presentes en la lista de abreviaturas del *Dicionário da Língua Portuguesa Contemporânea* de la Academia de Ciencias de Lisboa. Nuestro objetivo es vincular OntoDomLab-Med y las entradas seleccionadas del diccionario etiquetadas en TEI Lex-0, un sistema de marcado más estricto y más adecuado que la TEI para la codificación del diccionario, en línea con los principios FAIR. La ontología construida con Protégé y codificada en OWL permite la exportación de conocimiento en un formato de intercambio interoperable que permite que la ontología se aplique a diferentes recursos léxicos para hacer referencia a dominios independientemente del idioma utilizado.

OntoDomLab-Med será útil no solo para buscar información por dominio, sino que permitirá al lexicógrafo ser más consistente en su trabajo como investigador y codificador de información con fines lexicográficos.

Palabras clave: lexicografía, ontología de ámbito, marcas de ámbito, anotación lingüística, estándares

1. Introduction

Dans un monde globalisé tel que nous le connaissons aujourd'hui, les ressources lexicales – dictionnaires de langue, dictionnaires terminologiques, encyclopédies, vocabulaires, glossaires – sont un patrimoine linguistique et culturel essentiel dans une société multilingue, qui

occupent une place centrale dans les humanités numériques et le Web sémantique.

Les dictionnaires papier contemporains sont de plus en plus rares, alors que l'on retrouve les ressources lexicales numérisées un peu partout : elles sont aujourd'hui intégrées dans des sites Web, des applications mobiles et des services numériques. Ces ressources doivent être maintenues et requièrent une mise à jour permanente.

Les dictionnaires papier sont d'une richesse linguistique, patrimoniale et culturelle indéniables ; ils sont le résultat d'un travail lexicographique qui repose sur une compilation et description du lexique d'une langue en usage, confinée dans une macro et microstructure. Ce travail lexicographique repose sur une tradition de pratique de savoirs partagée par les lexicographes, plus que sur une méthodologie ancrée sur des critères explicites. Les informations lexicales et linguistiques contenues dans les dictionnaires font partie d'un patrimoine identitaire d'une communauté que nous avons le devoir de mettre à disposition du public sous un format recherchable en ligne.

La numérisation des dictionnaires s'impose. Transformer les dictionnaires papier existants en des ressources lexicales numériques est au centre de nos préoccupations. Nos axes de recherches tournent autour des méthodologies, des standards et des outils qui s'avèrent nécessaires pour structurer, hiérarchiser et annoter le lexique stocké dans les dictionnaires papier et les relier en les rendant consultables au même niveau que les ressources lexicales numérisées dès l'origine. L'objectif est de les rendre accessibles et interopérables sous un format normalisé pour les mettre en communication avec d'autres systèmes existants.

Une façon d'atteindre cet objectif est d'avoir recours à l'ontologie entendue comme une représentation computationnelle d'une conceptualisation (Gruber, 1995). Actuellement, le format le plus couramment utilisé pour la représentation d'ontologies est OWL[2] (Web Ontology Language), un format ouvert du Consortium W3C, qui est un langage de représentation des connaissances construit sur le modèle des données RDF[3] spécifiquement conçu pour le Web 2.0.

[2] www.w3.org/TR/owl2-primer/
[3] www.w3.org/RDF/

Dans le contexte de cet article, nous avons développé une ontologie des domaines appartenant aux sciences médicales et de la santé, OntoDomLab-Med. Nous partons d'une taxonomie élaborée à partir des domaines présents dans la liste des abréviations du *Dicionário da Língua Portuguesa Contemporânea* (DLPC) de l'Académie des Sciences de Lisbonne.

En amont, nous avons sélectionné quelques entrées qui contiennent les marques de domaine *Med.* [médecine] ou *Cirurg.* [chirurgie]. Ces marques ont été balisées en ayant recours au standard *de facto* TEI (Text Encoding Initiative), notamment TEI Lex-0 [4] qui vise à établir un encodage de base et un format cible pour faciliter l'interopérabilité des ressources lexicales annotées de façon hétérogène.

Mettre en lien OntoDomLab-Med[5] et des entrées sélectionnées du dictionnaire balisées en TEI Lex-0 est à la base de notre article. La méthodologie que nous préconisons est hybride, l'objectif étant d'associer une organisation des connaissances – une ontologie construite avec Protégé [6] – au traitement des données lexicographiques pour permettre le partage dans un format d'échange unique de façon à permettre une interopérabilité.

2. Restrictions d'usage dans les dictionnaires : les marques de domaines

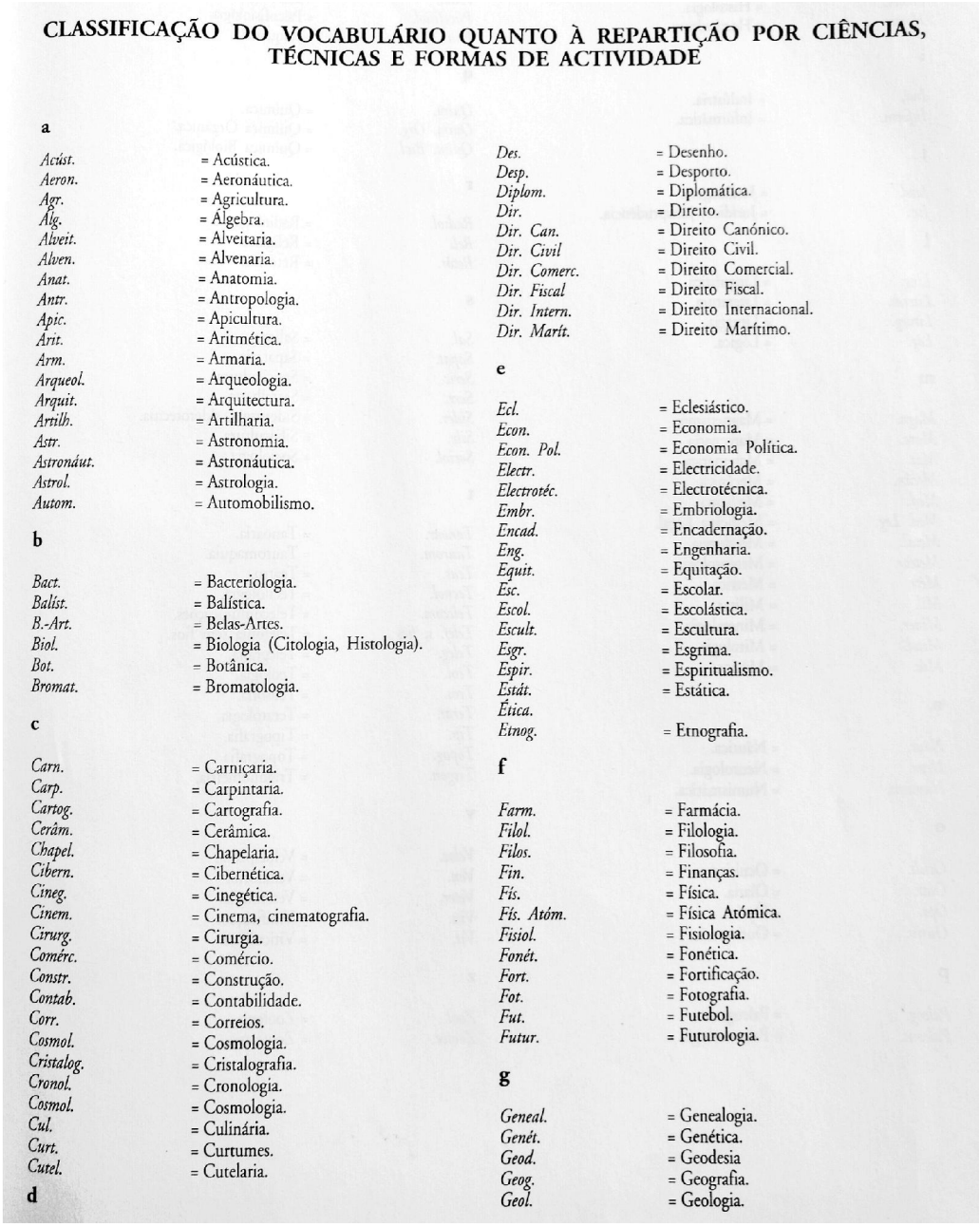
L'une des fonctions du dictionnaire de langue monolingue est de permettre la recherche de l'information concernant un mot : l'orthographe, le sens, l'étymologie, des synonymes, entre autres. D'après Rey (1982, 24), les dictionnaires *« doivent fractionner le discours selon un ordre résultant d'un classement ».* Ce classement est représenté dans la microstructure du dictionnaire par le biais des marques qui introduisent une organisation de l'information indiquant une restriction d'usage. Le système de marques lexicographiques (*labelling system*) représente l'une des thématiques les plus délicates des études métalexicographiques et de la lexicographie en général. La marque se définit comme étant un descripteur, généralement abrégé dans les

[4] https://dariah-eric.github.io/lexicalresources/pages/TEILex0/TEILex0.html
[5] https://github.com/sarampcarvalho/OntoDomLab-Med
[6] Un éditeur d'ontologies gratuit développé par Stanford Center for Biomedical Informatics Research : https://protege.stanford.edu.

dictionnaires papier et qui indique une restriction d'usage d'une unité lexicale par le biais d'une classification (domaine, registre…) (Salgado et al., 2020). Les mêmes auteurs, dans un autre article, définissent *marque de domaine* comme le *« marker that identifies the specialised field of knowledge in which a lexical unit is mainly used »* (Salgado et al., 2019a, 148).

Appliquer une marque d'usage à une unité lexicale signifie qu'elle s'éloigne *« in a certain respect from the main bulk of items described in a dictionary and that its use is subject to some kind of restriction »* (Svensén, 2009, 315). Le rôle des marques lexicographiques dans les dictionnaires est fondamental pour permettre à l'utilisateur un décodage de l'information qui est renfermée dans une microstructure bâtie par le lexicographe. Ces marques sont des dispositifs lexicographiques importants une fois qu'elles ont été choisies par les lexicographes pour bien servir les utilisateurs. La liste des désignations des marques que l'on retrouve dans les dictionnaires contiennent les abréviations qui habituellement sont organisées par ordre alphabétique, qui s'avère être un moyen efficace pour faciliter la localisation d'une marque donnée.

Dans le DLPC, les abréviations sont listées alphabétiquement dans une section appelée « Abreviaturas » [Abréviations] (DLPC, pp. XXXI-XXXIII), alors que les marques des domaines sont organisées dans une liste à part désignée « Classificação do vocabulário quanto à repartição por ciências, técnicas e formas de actividade » [Classification du vocabulaire quant à sa répartition par sciences, techniques et formes d'activités] (DPLC, pp. XXXV-XXXVI) :



CLASSIFICAÇÃO DO VOCABULÁRIO QUANTO À REPARTIÇÃO POR CIÊNCIAS, TÉCNICAS E FORMAS DE ACTIVIDADE

a

Acúst. = Acústica.
Aeron. = Aeronáutica.
Agr. = Agricultura.
Álg. = Álgebra.
Alveit. = Alveitaria.
Alven. = Alvenaria.
Anat. = Anatomia.
Antr. = Antropologia.
Apic. = Apicultura.
Arit. = Aritmética.
Arm. = Armaria.
Arqueol. = Arqueologia.
Arquit. = Arquitectura.
Artilh. = Artilharia.
Astr. = Astronomia.
Astronáut. = Astronáutica.
Astrol. = Astrologia.
Autom. = Automobilismo.

b

Bact. = Bacteriologia.
Balíst. = Balística.
B.-Art. = Belas-Artes.
Biol. = Biologia (Citologia, Histologia).
Bot. = Botânica.
Bromat. = Bromatologia.

c

Carn. = Carniçaria.
Carp. = Carpintaria.
Cartog. = Cartografia.
Cerâm. = Cerâmica.
Chapel. = Chapelaria.
Cibern. = Cibernética.
Cineg. = Cinegética.
Cinem. = Cinema, cinematografia.
Cirurg. = Cirurgia.
Comérc. = Comércio.
Constr. = Construção.
Contab. = Contabilidade.
Corr. = Correios.
Cosmol. = Cosmologia.
Cristalog. = Cristalografia.
Cronol. = Cronologia.
Cosmol. = Cosmologia.
Cul. = Culinária.
Curt. = Curtumes.
Cutel. = Cutelaria.

d

Des. = Desenho.
Desp. = Desporto.
Diplom. = Diplomática.
Dir. = Direito.
Dir. Can. = Direito Canónico.
Dir. Civil = Direito Civil.
Dir. Comerc. = Direito Comercial.
Dir. Fiscal = Direito Fiscal.
Dir. Intern. = Direito Internacional.
Dir. Marít. = Direito Marítimo.

e

Ecl. = Eclesiástico.
Econ. = Economia.
Econ. Pol. = Economia Política.
Electr. = Electricidade.
Electrotéc. = Electrotécnica.
Embr. = Embriologia.
Encad. = Encadernação.
Eng. = Engenharia.
Equit. = Equitação.
Esc. = Escolar.
Escol. = Escolástica.
Escult. = Escultura.
Esgr. = Esgrima.
Espir. = Espiritualismo.
Estát. = Estática.
Ética.
Etnog. = Etnografia.

f

Farm. = Farmácia.
Filol. = Filologia.
Filos. = Filosofia.
Fin. = Finanças.
Fís. = Física.
Fís. Atóm. = Física Atómica.
Fisiol. = Fisiologia.
Fonét. = Fonética.
Fort. = Fortificação.
Fot. = Fotografia.
Fut. = Futebol.
Futur. = Futurologia.

g

Geneal. = Genealogia.
Genét. = Genética.
Geod. = Geodesia
Geog. = Geografia.
Geol. = Geologia.

Figure 1 : Fragment de la liste de DLPC.

En général, les dictionnaires ne vont pas très loin dans l'information donnée aux utilisateurs ; ils se limitent à dire que toutes les marques évoquent des restrictions d'usage de l'information qui les suit :

endometrite [ẽdɔmɨtríti]. *s. f.* (De *endo-* + gr. μήτρα 'útero' + suf. *-ite*). *Med.* Inflamação da mucosa uterina.

Figure 2 : Entrée lexicographique « endometrite » [endométrite], DLPC.

À titre d'exemple, dans la Figure 2, la marque *Med.* introduit un texte définitoire, dont le sens est à interpréter à la lumière du domaine

médecine. Attribuer une marque à une unité lexicale donnée, la faire accompagner de la description du sens juste est une équation non pas sans risques. Une grande partie des unités lexicales est multi-domaines, c'est-à-dire qu'elles sont employées dans une grande diversité d'usages de pratiques et de communications. Décider lequel des sens décrire pour une marque attribuée à une unité lexicale spécialisée ressort de la responsabilité de l'équipe de lexicographes qui généralement prend sa décision sur la base d'une expérience acquise, par tradition ou par mimétisme.

Les lexicographes sont confrontés à de multiples problèmes liés à la manière d'enregistrer ces informations et aux limites du système de marques lexicographiques. Ptaszyński (2010, 411) étudie les causes d'un traitement théorique insatisfaisant des informations diasystématiques dans les dictionnaires et conclut que les lexicographes *« have been searching in vain for an exhaustive and precise answer to the questions of which words to label in what kind of dictionaries and how to do it »*. Selon le même auteur, ces problèmes résultent d'un *« lack of a firm theoretical basis for the application of diasystematic information (i. e. information about restrictions on usage) in dictionaries »* (Ibid.).

Atkins et Rundell (2008, 496) sont conscients de la difficulté de la tâche quand ils affirment *« labelling is an area of lexicography where there is more work to be done »*. À Atkins et Rundell s'associent des auteurs tels que Svensén (2009), Bergenholtz et Tarp (1995) et Hausmann (1989) qui considèrent que l'on est encore loin d'arriver à établir des critères solides de balisages qui permettraient aux lexicographes d'annoter leurs données de façon systématique et satisfaisante. Atkins et Rundell (2008, 231) affirment que *« There's quite a lot of work involved in putting together a consistent policy on labels in a dictionary »,* tandis que Sakwa (2011, 308) déclare que *« there is no agreed-on criteria for making usage decisions »,* et Fedorova (2004, 265) considère que *« there is no consistency in the labelling policy »*.

Si Béjoint (2010), Atkins et Rundell (2008) et Landau (2001), entre autres, défendent qu'une théorie de la lexicographie n'existe pas, cela ne veut pas dire que les principes qui guident les lexicographes ne doivent pas reposer sur des théories linguistiques (cf. Atkins et Rundell, 2008, 4-10). Nous pensons, qu'en plus des théories linguistiques, les *« principles that guide lexicographers »* défendus par Atkins et Rundel (2008, 9) doivent inclure des méthodologies d'aide à la décision, en plus des standards et

des outils qui permettent d'avancer dans l'analyse des données lexicographiques et leurs représentations en tenant compte aussi des principes FAIR [Findable, Accessible, Interoperable, Reusable].[7]

C'est bien en ligne avec les principes FAIR que nous concevons l'utilisation des ontologies et de TEI Lex-0 pour traiter les marques de domaines dans le DLPC.

3. Ontologies et interopérabilité sémantique

3.1 L'interopérabilité sémantique

L'une des thématiques centrales de notre article est l'interopérabilité, que l'Institut des ingénieurs électriciens et électroniciens (IEEE) a définie comme étant *« the ability of two or more systems or components to exchange information and to use the information that has been exchanged »* (Geraci, 1991, 42). Dans le domaine des technologies de l'information, cette définition a été révisée et mise à jour. La norme ISO / IEC 2382 : 2015 définit l'interopérabilité comme suit : *« capability to communicate, execute programs, or transfer data among various functional units in a manner that requires the user to have little or no knowledge of the unique characteristics of those units ».* Deux points pertinents sont communs aux définitions du concept d'interopérabilité ici présenté. Les informations : (i) sont échangées entre deux ou plusieurs systèmes et, (ii) doivent être utilisables et, *a fortiori*, réutilisables. Cet échange d'informations se construit à plusieurs niveaux. Le premier niveau identifié se situe au plan de l'infrastructure technique (par exemple, les réseaux et les protocoles) qui permet que l'échange des données s'accomplisse. Le deuxième niveau se situe au plan de l'interopérabilité syntaxique qui nous permet d'avoir accès à des formats de données communs et, avant tout, à leur couche sémantique. C'est ainsi que l'échange des informations s'appuie sur une compréhension partagée non seulement de la *structure* de ce qui est transmis, mais aussi de la *signification* du message (interopérabilité sémantique), dont le contenu est défini sans ambiguïté.[8]

Des solutions technologiques sont proposées qui pour la plupart correspondent au paradigme du XXI[e] siècle du Web sémantique : un Web de données dans lequel les machines peuvent communiquer plus

[7] www.go-fair.org/fair-principles/
[8] www.w3.org/2001/sw/BestPractices/OEP/SemInt/

efficacement, car l'information *« is given well-defined meaning »* (Berners-Lee, Hendler et Lassila, 2001, 35).

3.2 Ontologie et ontologies

Notre article se focalise sur l'importance du rôle des ontologies pour la mise en œuvre de l'interopérabilité sémantique dans l'ensemble des technologies qui englobe la Web sémantique.[9]

Parler de la notion d'ontologie implique choisir entre deux significations possibles. Lorsque « Ontologie » est écrite en majuscule, elle fait référence à une branche de la philosophie qui se concentre sur l'étude de l'être, de ce qui est (être en tant qu'être), comme l'a affirmé Aristote dans *Catégories*[10]. Sur la base de l'étude de l'être, Aristote propose dix catégories ontologiques, à savoir Substance, Quantité, Qualité, Relation, Lieu, Temps, Position, État, Action et Affection, visant à rendre compte des différents types de connaissances dans le monde. Pour compléter cette idée, Aristote propose une classification des prédicats, à savoir, la propriété, la définition (y compris la *differentia*), le *genus* et l'accident (*Topics*) ultérieurement étendue par le philosophe Porphyre dans son œuvre fondatrice *Isagoge*, qui a servi d'introduction aux *Catégories*. Dans *Isagoge*, Porphyre a distingué cinq types de prédicats – *genus*, espèce, *differentia*, propriété et accident – et a fourni un exemple remarquable de division de substance par une différence spécifique, représenté visuellement dans ce qui est connu aujourd'hui comme étant l'arbre de Porphyre :

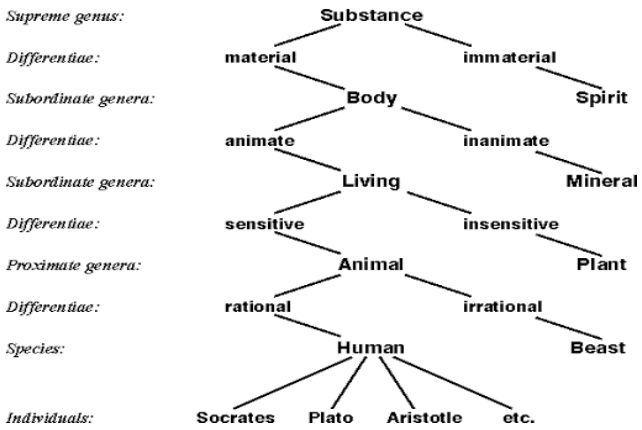


Figure 3 : Arbre de Porphyre (Isagoge).

[9] https://obitko.com/tutorials/ontologies-semantic-web/semantic-web-architecture.html
[10] www.perseus.tufts.edu/hopper/text?doc=Perseus:abo:tlg,0086,025:4

Cette représentation considérée comme l'une des plus influentes jusqu'à nos jours illustre la subdivision continue des différents *genera* jusqu'à arriver aux espèces les plus spécifiques. L'influence de l'arbre de Porphyre pour les ontologies modernes est également présente dans le principe d'hérédité, c'est-à-dire dans le fait que l'espèce hérite les *differentiae* des respectifs *genera*.

Depuis la fin du XX[e] siècle, en particulier dans les années 80, les ontologies (cette fois en minuscules et au pluriel) (cf. Guarino et Giaretta, 1995) ont assumé une nouvelle fonction et une nouvelle ampleur dans les domaines de l'informatique, de l'intelligence artificielle et de l'ingénierie des connaissances. Selon ces perspectives, les ontologies sont considérées comme des artefacts computationnels utilisés pour *« formally model the structure of a system »* (Guarino, Oberle et Staab, 2009, 2), notamment à travers une conceptualisation, c'est-à-dire à travers la représentation des types d'entités pertinents dans un domaine donné, ainsi que les relations entre elles.

La définition des ontologies la plus populaire à cet égard est sans doute celle de Gruber (1995, 908), qui définit l'ontologie comme étant *« an explicit specification of a conceptualization »*. À cette définition il ajoute l'idée que dans les *« knowledge-based systems, what 'exists' is exactly that which can be represented »* (*ibid.*). Borst (1997) a ensuite retravaillé la proposition de Gruber et a introduit les notions de « spécification formelle » et de « conceptualisation partagée ». Un an plus tard, Studer et al. (1998) ont fusionné les définitions précédentes et en ont proposé une nouvelle. Pour eux, une ontologie est *« a formal, explicit specification of a shared conceptualization »* (185). C'est cette dernière définition qui est utilisée dans le cadre de ce travail, puisqu'elle introduit trois caractéristiques centrales : premièrement, cette spécification doit être explicite, c'est-à-dire que le type de concepts et les contraintes respectives sur leurs usages doivent être explicitement définis ; deuxièmement, elle doit être formelle, c'est-à-dire compréhensible pour une machine ; et troisièmement, elle doit être partageable, parce qu'elle doit être en mesure de saisir les connaissances qui sont consensuelles pour une communauté donnée.

Plusieurs propositions concernant la classification des ontologies ont été faites par différents auteurs. En ce qui concerne leur structure, Guarino (1998) a distingué les « ontologies de niveau supérieur », qui visent à décrire

les catégories indépendantes du domaine, les « ontologies de domaine », qui décrivent les concepts fondamentaux d'un domaine donné, les « ontologies de tâches », qui décrivent les concepts se rapportant à une tâche spécifique et, finalement, les « ontologies d'applications », qui correspondent à d'autres ontologies spécialisées, conçues pour un cas d'utilisation ou une application particulière. Cette dernière catégorie englobe l'ontologie proposée dans cet article. Selon Arp, Smith et Spear (2015), les ontologies d'applications sont construites avec un objectif précis et, par conséquent, leurs portées répondent aux contraintes et aux objectifs inhérents au projet en question.

3.3 Opérationnalisation et Web Ontology Language (OWL)

Les avantages des ontologies en matière de représentation et de gestion des connaissances ont été mis en avant par Uschold et Gruninger (1996) : une fois utilisées et partagées pour des applications et des scénarios divers, les ontologies permettent l'interopérabilité entre systèmes. En permettant un contrôle automatique de la consistance, les ontologies conduisent à des produits et des systèmes plus fiables. Et finalement, en visant à définir explicitement l'ensemble des concepts dans un domaine spécifique sur la base d'une compréhension partagée d'une réalité donnée, les ontologies fournissent un cadre utile pour les individus d'horizons et de perspectives différents par rapport à cette réalité.

Pour rendre cette opérationnalisation possible dans le cadre du Web sémantique, les ontologies ont recours au *Web Ontology Languag*e (OWL), un langage de calcul développé par le World Wide Web Consortium (W3C) et intégré dans la technologie du Web sémantique. Conçu pour *« represent rich and complex knowledge about things, groups of things, and relations between things »*[11], OWL recourt à des axiomes (des énoncés de base exprimés par une ontologie OWL), à des entités (classes, individus, propriétés) et à des contraintes pour construire une représentation logique axée sur un domaine donné, renforcée par la puissance des raisonneurs, qui vérifient la cohérence des connaissances exprimées et rendent explicites les connaissances implicites à travers l'inférence. Le potentiel de l'interopérabilité des ontologies s'est amélioré dans la version actuelle du

[11] www.w3.org/TR/owl2-primer/

langage OWL 2 [12], grâce à l'utilisation des IRI [13] (Internationalized Resource Identifiers) pour identifier à la fois les ontologies elles-mêmes[14] et leurs éléments respectifs. Ces identificateurs seront particulièrement utiles pour le travail présenté ici, car les IRI constituent une passerelle convenable pour relier les composantes ontologiques pertinentes aux entrées du dictionnaire, permettant ainsi l'interopérabilité sémantique.

4. Les marques de domaines dans le DLPC

Le *Dicionário da Língua Portuguesa Contemporânea* (DLPC) est un dictionnaire monolingue de langue portugaise publié en 2001 par l'Académie des Sciences de Lisbonne, avec le soutien financier de la Fondation Calouste Gulbenkian, sous la responsabilité commerciale de l'éditeur Verbo. Ce travail lexicographique a été dirigé par Malaca Casteleiro et réalisé par une grande équipe de collaborateurs. Le DLPC contient environ 70.000 entrées et accumule 167.556 sens. L'édition imprimée comprend deux volumes, totalisant 3.880 pages.

Après une longue pause, le travail a été repris en 2015. Aujourd'hui la version PDF du DPLC a été convertie en XML à l'aide d'une version personnalisée du schéma P5 de TEI. Le DLPC est actuellement en cours de conversion au format TEI Lex-0 à des fins d'interopérabilité des données (Salgado et al., 2019b). Cette version préliminaire du dictionnaire est disponible sous forme de base de données et constitue le point de cette recherche en cours.[15]

Le DLPC, comme tout dictionnaire académique, fait autorité. Ce fait le rend particulièrement intéressant à étudier, car les résultats obtenus ont de l'impact dans la communauté de lexicographes et les utilisateurs

[12] Lancé 2009 et mise à jour en 2012 pour remplacer OWL 1 (2004).

[13] www.w3.org/International/O-URL-and-ident.html

[14] Chaque ontologie peut avoir un IRI, qui est utilisé pour identifier la ressource elle-même. En complément, il peut y avoir une version IRI pour identifier l'actuelle version de l'ontologie. La version IRI peut, mais pas forcément, correspondre à IRI de l'ontologie (www.w3.org/TR/owl2-syntax/#IRIs).

[15] La nouvelle édition numérique n'est pas publique. Ana Salgado (ACL, NOVA CLUNL) est la coordinatrice de la nouvelle édition numérique en cours. L'équipe est constituée par Alberto Simões (IPCA), José João Almeida et Álvaro Iriarte Sanromán (tous les deux de Université Minho). La participation de NOVA CLUNL est liée à la transition du dictionnaire au format TEI Lex-0 avec la participation de Toma Tasovac (BCDH) et Ana Salgado et au développement de l'ontologie avec Sara Carvalho et Rute Costa.

peuvent bénéficier des bonnes pratiques qui dérivent, nous l'espérons, de la méthodologie appliquée. Étant donné que notre but est de construire une ontologie des domaines, le DLPC s'avère être un excellent dictionnaire car il couvre au total 184 domaines, qui vont de la médecine à la chapellerie. Les analyses qui ont été faites au DLPC nous montrent que le marquage des domaines et sous-domaines n'a pas été fait sur la base de critères clairs, pour la bonne raison que nous n'arrivons pas à identifier une cohérence dans la pratique de marquage utilisée. Néanmoins, le DLPC est un dictionnaire savant, parce qu'il s'inclut typiquement dans le groupe des dictionnaires *« traditionally designed with such main features as the pursuit of completeness with regard to the entries relevant to subject matters »* (Kinable, 2015, 12). Les informations lexicographiques détaillées et une microstructure d'un niveau de complexité élevé posent des défis intéressants pour une modélisation cohérente des données.

Notre choix étant fait, nous sommes passés à la requête et à l'analyse de toutes les marques de domaines du DPLC. Ce travail avait déjà au préalable été fait manuellement à partir de la liste des abréviations du dictionnaire (Salgado et Costa, 2019). Nous sommes ensuite passés au dictionnaire disponible en base de données, et nous avons pu constater que des 184 domaines identifiés, quoique répertoriés dans la liste des abréviations, il y en avait qui n'apparaissaient pas dans le dictionnaire, tels que « Bromatologia » [bromatologie], « Cibernética » [cybernétique], « Economia Política » [économie politique], « Escolástica » [scolastique], « Espiritualismo » [spiritualisme], « Futurologia » [futurologie], « Polícia » [police], « Química Biológica » [chimie biologique], « Química Orgânica » [chimie organique], « Telefonia sem Fio » [téléphonie sans fil] et « Velocipedia » [velocipedia].

C'est la raison pour laquelle des 184 domaines identifiés, nous avons soustrait les 11 domaines qui n'ont pas de registre dans le DLPC et nous sommes passés à 173 domaines. D'autre part, plusieurs marques qui identifient des lemmes dans le dictionnaire, mais qui ne figurent pas dans la liste des abréviations ont été détectées. Nous avons pu les repérer dans la base de données. C'est le cas de « Bioquímica » [biochimie], « Etnologia » [ethnologie], « Metrologia » [métrologie], « Agronomia » [agronomie], « Marítima » [maritime], « Psycanálise » [psychanalyse], « Ecologia » [écologie], « Ginástica » [gymnastique], « História Política »

[histoire politique], « Pirotecnia » [pyrotechnie] et « Transportes » [transports]. Nous avons donc décidé de les ajouter à la liste des 173 domaines et nous avons à nouveau obtenu 184 domaines :

DLPC Domaines	#	DLPC Domaines	#	DLPC Domaines	#
Botânica	3494	Pintura	92	Serralharia	8
Zoologia	3203	Numismática	91	Silvicultura	8
Medicina	2430	Equitação	84	Tanoaria	8
Religião	1489	Aeronáutica	79	Trigonometria	8
Náutica	1397	Óptica	77	Aritmética	7
Química	1371	Carpintaria	69	Balística	7
Física	1110	Artilharia	64	Filologia	7
Música	1057	Metalurgia	64	Genealogia	6
Militar	890	Mitologia	64	Neurologia	6
Linguística	848	Automobilismo	60	Olaria	6
História	823	Métrica	56	Psicofisiologia	6
Anatomia	816	Antropologia	54	Sapataria	6
Jurídico, jurisprudência	785	Vinificação	53	Cronologia	5
Política	718	Fortificação	51	Hidrografia	5
Biologia (citologia, histologia)	706	Encadernação	48	Jardinagem	5
Filosofia	667	Contabilidade	47	Paleografia	5
Matemática	544	Etnografia	47	Pecuária	5
Direito	517	Viticultura	47	Piscicultura	5
Culinária	513	Histologia	43	Teratologia	5
Desporto	505	Pedagogia	41	Cartografia	4
Gramática	495	Telecomunicações	41	Curtumes	4
Geologia	471	Direito Internacional	40	Etnologia	4
Literatura	436	Escolar	39	Geodesia	4
Agricultura	417	Arqueologia	38	Higiene	4
Astronomia	413	Direito Canónico	36	Metrologia	4
Arquitectura	410	Paleontologia	32	Agronomia	3
Tipografia	369	Astronáutica	30	Cerâmica	3
Geometria	359	Astrologia	29	Marcenaria	3
Economia	358	Electrotécnica	29	Marinha	3
Construção	340	Cristalografia	28	Medicina Legal	3
Psicologia	332	Topografia	28	Psicanálise	3
Geografia	267	Zootecnia	26	Siderurgia, siderotecnia	3
Belas-Artes	266	Direito Comercial	25	Teosofia	3
Mineralogia	254	Engenharia	24	Venatório	3
Farmácia	235	Bacteriologia	22	Ecologia	2
Teatro	234	Ocultismo	21	Estática	2
Fisiologia	212	Escultura	18	História Natural	2
Pesca, pescaria	206	Salinas	18	Horticultura	2
Informática	201	Carniçaria	17	Pré-história	2
Tecnologia	198	Acústica	16	Alveitaria	1
Fonética	180	Apicultura	15	Alvenaria	1
Psiquiatria	177	Cinegética	15	Cutelaria	1
Mecânica	163	Direito Civil	14	Ginástica	1
Veterinária	163	Ourivesaria	14	História Política	1
Liturgia	156	Álgebra	13	Magnetismo	1
Eclesiástico	152	Correios	13	Pirotecnia	1
Electricidade	151	Genética	13	Radiologia	1
Meteorologia	148	Telegrafia	13	Sericicultura	1
Tauromaquia	146	Desenho	12	Transportes	1
Teologia	145	Esgrima	11	Bromatologia	0
Retórica	142	Chapelaria	10	Cibernética	0
Heráldica	139	Cosmologia	10	Economia Política	0
Armaria	132	Ética	10	Escolástica	0
Cinema, cinematografia	129	Física Atómica	10	Espiritualismo	0
Fotografia	124	Patologia	10	Futurologia	0
Cirurgia	120	Diplomática	9	Polícia	0
Futebol	120	Direito Fiscal	9	Química biológica	0
Indústria	110	Direito Marítimo	9	Química orgânica	0
Comércio	105	Hidráulica	9	Telefonia sem fios	0
Lógica	99	Bioquímica	8	Velocipedia	0
Finanças	94	Embriologia	8		
Sociologia	94	Parasitologia	8		

Figure 4 : Domaines et occurrences trouvés dans le DLPC

Après avoir analysé les domaines listés (Figure 4), nous avons détecté des inconsistances dans les domaines. Dans cette liste, nous trouvons des domaines qui se trouvent au même niveau que leurs sous-domaines. À titre d'exemple, le domaine des mathématiques a le même statut que l'algèbre, l'arithmétique, la géométrie et la trigonométrie. Du point de vue de la consistance, de l'organisation de l'information du dictionnaire et surtout d'un point de vue ontologique la coexistence nivelée de domaines et sous-domaines nous semble un mauvais choix, car cela ne correspond pas à la façon dont les sciences ou les activités s'organisent. Qu'est-ce qui a justifié ce choix ?

Nous allons nous attarder plus longuement sur les domaines et sous-domaines qui concernent la médecine. Le déséquilibre entre domaines et sous-domaines que nous venons de mentionner se reflète aussi dans ce cas. Nous pouvons trouver, dans la liste du DPLC, le domaine plus général « Medicina » [médecine] (2430) qui cohabite avec des spécialités médicales tels que « Cirurgia » [chirurgie] (120), « Medicina Legal » [médecine légale] (3), « Neurologia » [neurologie] (6), « Patologia » [pathologie] (10), « Psiquiatria » [psychiatrie] (177), « radiologie » [radiologie] (1).

Le nombre réduit d'entrées associées à la médecine légale (3), à la neurologie (6) ou à la radiologie (1) est surprenant. En regard des fréquences réduites, nous pouvons mettre en cause la pertinence de l'existence de ces marques dans le dictionnaire. Pourtant, nous sommes en face d'un problème usité dans la pratique lexicographique. L'attribution de la marque est généralement effectuée par le lexicographe qui n'a d'autre instrument pour prendre des décisions que son intuition.

En amont, le DLPC contient des marques de domaines qui, selon le point de vue, peuvent appartenir au domaine de la médecine ou être des domaines à part entière. C'est le cas de l'anatomie : si on considère l'anatomie pathologique, on la classe dans le domaine de la médecine ; si on la considère comme anatomie générale, on la classe dans le domaine de la biologie. Toujours selon le même raisonnement, la marque « Tératologie » [tératologie] (5) est directement liée à l'Anatomie pathologique. Pourquoi lui avoir donné le statut de domaine alors qu'elle n'a que 5 occurrences ? Que faire de « Bacteriologia » [bactériologie], « Genética » [génétique] (13), « Embriologia » [embryologie] (8),

« Fisiologia » [physiologie] (212), « Histologia » [histologie], ou « Parasitologia » [parasitologie] (8) ? Toutes ces marques de domaines posent des problèmes.

Pour illustrer notre méthodologie, nous avons choisi toutes les marques de domaines liées à la Médecine. Ce choix est dû à plusieurs raisons. Parler de marques de domaines implique parler d'unités lexicales qui sont des unités employées pour désigner et dénommer des concepts de domaines qui en langue acquièrent des sens qui sont décrits par les lexicographes dans les dictionnaires de langue générale. Les lexicographes connaissent les sens des unités lexicales, mais ils ont plus de difficulté à décrire les sens des unités lexicales spécialisées. C'est pour aider les lexicographes dans leur tâche que nous proposons la construction d'une ontologie.

5. Ontologie d'application pour le DLPC : OntoDomLab-Med

Cette section de l'article vise à décrire OntoDomLab-Med, l'ontologie d'application que nous avons développée pour les sciences médicales et de la santé. OntoDomLab-Med sert non seulement à améliorer la cohérence dans l'attribution des marques de domaine à des entrées lexicographiques spécifiques, mais elle permet aussi de faire la liaison avec les données lexicographiques annotées avec TEI Lex-0 et assurer une opérationnalisation sémantiquement plus précise pour l'échange des informations.

En ayant comme point de départ la taxonomie des marques des domaines du DLPC, nous l'avons ensuite comparée à d'autres systèmes de classifications de marques de domaines existants, telles que la Classification décimale universelle (CDU) [16], le thésaurus de l'UNESCO[17], les WordNet Domains[18] et EuroSciVoc[19]. La Modern Science Ontology (ModSci)[20], *« an upper ontology that provides a unifying framework for the various domain ontologies that make up the Science Knowledge Graph Ontology Suite »,* a également été analysée. Compte tenu de la décision d'élaborer une ontologie des sciences médicales et de la santé,

[16] www.udcsummary.info/php/index.php
[17] http://vocabularies.unesco.org/browser/thesaurus/fr/
[18] http://wndomains.fbk.eu/
[19] https://op.europa.eu/en/web/eu-vocabularies/euroscivoc
[20] https://saidfathalla.github.io/Science-knowledge-graph-ontologies/doc/ModSci_doc/index-en.html#

nous avons aussi procédé à une analyse des ressources davantage axées sur les soins de la santé, à savoir SNOMED CT[21], MeSH[22], Unified Medical Language System (UMLS)[23] et Disease Ontology.[24]

OntoDomLab-Med a été construit avec Protégé, qui est actuellement l'éditeur OWL le plus utilisé. Son succès est dû au fait qu'il s'agit d'un environnement d'édition d'ontologies riche en fonctionnalités, parfaitement adapté à OWL 2 et qui permet des connexions directes avec des raisonneurs de logique de description tels que HermiT[25] (développé par l'Oxford University Computing Laboratory).

Pour construire OntoDomLab-Med nous avons défini deux prémisses : *(i)* concevoir une hiérarchie des classes solide *(ii)* qui de surcroît pourrait soutenir des relations non-hiérarchiques et ainsi favoriser une force expressive au sein de l'ontologie. Lorsque l'on élabore une hiérarchie des classes fondamentales dans une ontologie, surtout si la source principale est proche de la structure d'un thesaurus, comme c'est le cas pour la plupart des ressources consultées, nous ne pouvons pas négliger que la relation *broader* terme / *narrower* terme sous-jacent à l'organisation taxonomique ne correspond pas nécessairement à la relation de subsomption transitive `is_a`. (Doerr, 2001 ; Thiéblin et al., 2017). Comme l'ont affirmé Baker et al. (2013, 37) :

> *[...] informally defined KOSs cannot typically be translated into the language of RDFS and OWL properties and classes, with their formal-logical implications, without introducing potentially false or misleading logical precision. Informal KOSs may be converted into formal ontologies [...], but the process of assigning appropriate formal semantics to the elements of a KOS may require a long, hard modeling effort. Hierarchical relationships, for example, must be disambiguated into relationships of class instantiation, class subsumption, part–whole, or other types—a process that cannot usually be automated.*

Par exemple, l'unité « anatomie » peut avoir « médecine » comme un *narrower* term, mais cela ne veut pas dire que « anatomie » est une sous-classe de `médecine`. La relation à médecine peut être étayée si l'unité

[21] http://bioportal.bioontology.org/ontologies/SNOMEDCT
[22] http://bioportal.bioontology.org/ontologies/MESH
[23] www.nlm.nih.gov/research/umls/index.html
[24] http://bioportal.bioontology.org/ontologies/DOID
[25] www.hermit-reasoner.com/

est étendue et spécifiée : « pathologie anatomique ». Autrement, « anatomie » est un domaine de la « biologie ».

En 2007, afin de permettre aux différents systèmes d'organisation des connaissances – tels que les schémas de classification, les systèmes de vedettes-matières, les thésaurus ou les taxonomies – d'être exprimés comme des données lisibles par la machine à travers le Web sémantique, SKOS[26] (Simple Knowledge Organization System) a été proposé comme une recommandation du W3C. Cependant, en tant que modèle de données, sa portée diffère grandement de celle de OWL dans le sens où les relations `skos:broader` et `skos:narrower` ne sont pas transitives, c'est-à-dire, qu'indiquer que A `skos:broader` B et B `skos:broader` C n'implique pas que A `skos:broader` C. À cet égard, Derriere et al. (2009, 4) déclare catégoriquement que *"a vocabulary (SKOS or otherwise) is not an ontology. It has lighter and looser semantics than an ontology and is specialised for the restricted case of resource retrieval"*.

Dans le cadre d'OntoDomLab-Med, nous avons décidé de ne pas transposer directement les domaines originaux de la taxonomie de DLPC en tant que hiérarchie orientée vers une subsomption, afin d'éviter les représentations trompeuses. Au lieu de cela, la relation `hasBranch` et sa relation inverse `BranchOf` ont été ajoutées à la hiérarchie des propriétés d'objet (Object Property dans Protégé) pour soutenir cette proposition d'organisation des connaissances et faciliter la liaison subséquente avec les informations lexicographiques annotées en TEI Lex-0.

En haut de la hiérarchie principale, `MedicalAndHealthSciences` est représenté comme un concept de base[27] comme suit :

```
MedicalAndHealthSciences is_a FieldOfScience
MedicalAndHealthSciences hasBranch Medicine
MedicalAndHealthSciences hasBranch HealthSciences
```

Outre des marques des domaines liées à la médecine recueillies à l'origine dans le DLPC d'autres ont été ajoutées à l'ontologie, ce qui fait

[26] www.w3.org/2004/02/skos/

[27] `MedicalAndHealthSciences` provient de la nomenclature proposée dans la ressource EuroSciVoc et permet à OntoDomLab-Med de se développer à l'avenir, en intégrant non seulement les concepts liés à la médecine, mais aussi de lier ceux qui appartiennent aux sciences de la santé.

un total de 19 sous-domaines (Figure 5), représentés en OntoGraf[28] plug-in de Protégé.

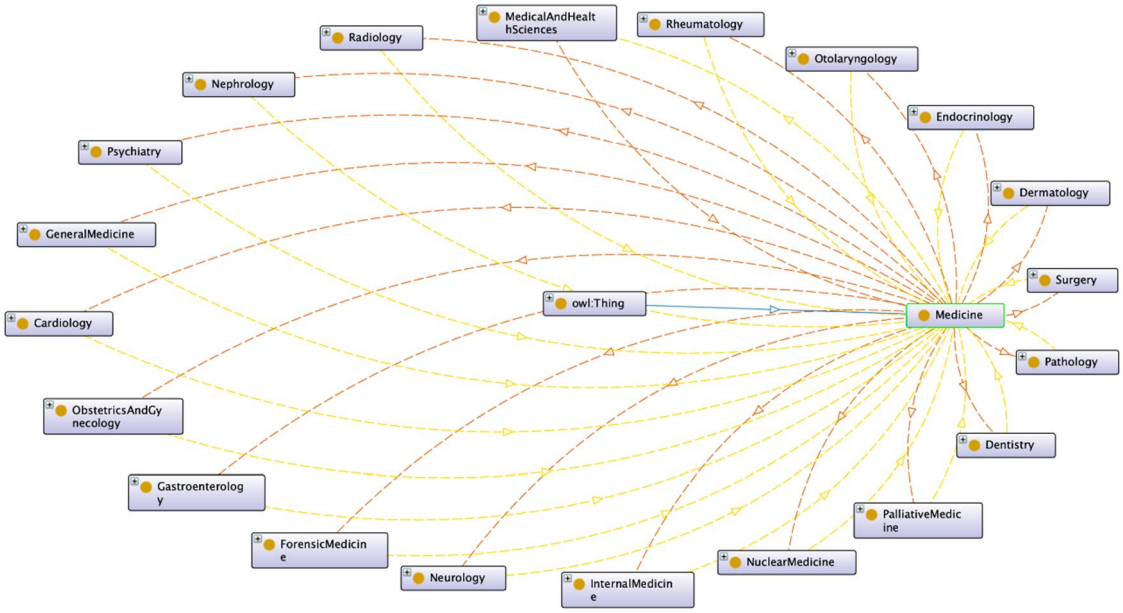


Figure 5 : Représentation de la hiérarchie noyau de OntoDomLab-Med avec OntoGraf.

`NaturalSciences` et ses domaines correspondants, à savoir la biologie, ont également été ajoutés à OntoDomLab-Med pour illustrer la capacité de l'ontologie à prendre en charge la polyhiérarchie, par opposition aux représentations taxonomiques mono-hiérarchiques typiques. Cette fonctionnalité fournit une valeur ajoutée pour l'annotation des domaines dans les éditions du DLPC à venir, permettant un marquage plus fin, et de ce fait permettre la désambiguïsation dans des scénarios particuliers. L'anatomie représente un exemple concret de tels avantages : la marque *Anat.* est actuellement indistinctement attribuée à des entrées liées soit à la médecine, soit à la biologie, ce qui conduit à une ambiguïté potentielle. Avec l'appui de l'ontologie, une distinction entre la marque de domaine Anatomy `BranchOf` Biology et celle de `AnatomicPathology` `BranchOf` Medicine peut être faite.

Afin d'améliorer la puissance expressive de OntoDomLab-Med, notamment en tirant parti des capacités de validation logique liées à OWL 2, un ensemble de 10 relations non hiérarchiques (y compris `hasBranch` et `BranchOf`), ainsi que leurs relations inverses respectives, ont

[28] https://protegewiki.stanford.edu/wiki/OntoGraf

été ajoutées à la hiérarchie des Object Property de l'ontologie. L'ensemble complet des relations étendues est présenté dans la Table 1.

RELATION	RELATION INVERSE
hasAssociatedMorphology	AssociatedMorphologyOf
hasBranch	BranchOf
hasDirectMorphology	DirectMorphologyOf
hasDirectProcedureSite	DirectProcedureSiteOf
hasFinding	FindingOf
hasFindingSite	FindingSiteOf
hasMethod	MethodOf
hasProcedure	ProcedureOf
treats	isTreatedBy
usesDevice	DeviceOf

Table 1 : Ensemble de relations non-hiérarchiques ajouté à OntoDomLab-Med.

Le but a donc été de permettre une représentation détaillée des connaissances pouvant aller au-delà des spécialités médicales décrites dans la hiérarchie principale, en reliant ces dernières à des concepts d'autres domaines connexes (par exemple, une spécialité médicale traite une certaine maladie, qui a un ensemble de symptômes et peut affecter une ou plusieurs structures anatomiques). Cela contribuera à son tour à établir des définitions formelles de concept, qui sont utiles pour rédiger des définitions en langue naturelle (Carvalho et al., 2018). Ainsi, pour mieux faire usage de l'extension d'OntoDomLab-Med et obtenir des définitions formelles de concept interopérables, les hiérarchies suivantes ont été incluses dans l'ontologie :

- `Action` (fait référence à l'action noyau qui a sous-jacente une procédure clinique)
- `AnatomicalEntity` (fait référence à la structure anatomique qui a été affectée par une maladie ou une procédure)
- `ClinicalFinding` (fait référence aux signes et symptômes qui caractérisent une maladie donnée)

- `Device` (utilisé dans une procédure clinique)
- `Disease` (fait référence à toute maladie ou trouble qui affectent l'être humain)
- `MorphologicAbnormality` (fait référence à tout type de lésion)
- `PathologicalProcess` (fait référence à tout procès qui est nuisible, c'est-à-dire, qui entraîne une condition pathologique)
- `Procedure` (se réfère à toute procédure clinique, qui peut être chirurgical ou non-chirurgicale, invasive ou non-invasive, etc.)

De plus, OntoDomLab-Med a eu recours à des fonctionnalités de OWL 2 telles que les contraintes de domaine et de *range* (c'est-à-dire, les entités auxquelles s'applique une propriété déterminée), les restrictions de propriété (à savoir la propriété inverse, ainsi que la symétrique et l'asymétrique) (cf. Figure 6), et la disjonction des classes (une sorte de « relation d'incompatibilité » impliquant 2 classes dans l'ontologie, ce qui veut dire qu'aucun individu ne peut être une instance des deux classes en question), afin de bénéficier pleinement des capacités de vérification logique et d'inférence du raisonneur.

Figure 6 : Exemple de relations non-hiérarchiques dans OntoDomLab-Med avec leurs relations inverses, ainsi que les restrictions qui concernent le domaine et range.

L'exemple suivant (Figure 7) illustre la définition formelle du concept <endométrite> dans OntoDomLab-Med.

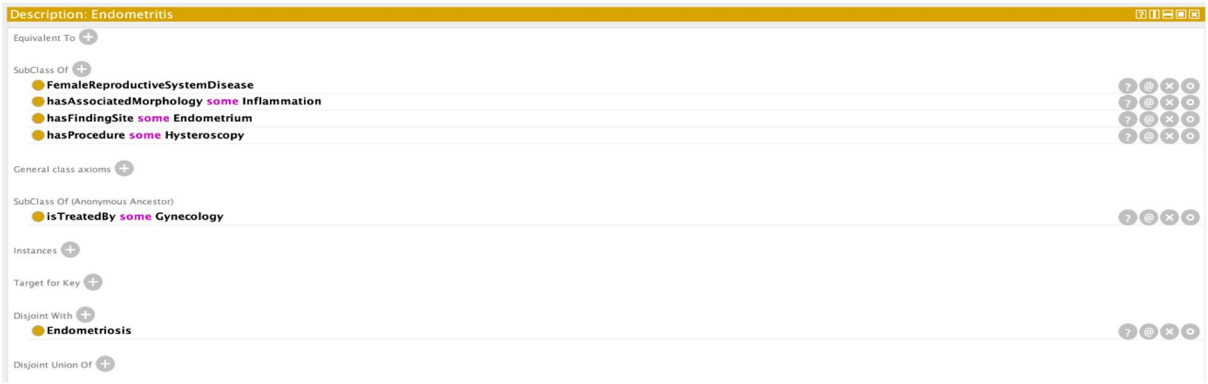


Figure 7 : Définition formelle du concept de <endométrite>.

Comme nous l'avons mentionné préalablement, chaque composante de l'ontologie a un IRI qui lui est associé, ce qui permet la liaison avec TEI. Dans ce cas, l'IRI pour l'endométrite serait :

`www.semanticweb.org/OntoDomLab-Med#Endometritis`

En plus, cette représentation ontologique permet une navigation par concept, ce qui signifie que l'on peut accéder à la définition formelle du concept de <hystéroscopie>, une procédure non chirurgicale associée au diagnostic et au traitement de l'endométrite. La Figure 8 illustre la définition formelle de la procédure, lisible par machine.

```
<!-- http://www.semanticweb.org/OntoDomLab-Med#Hysteroscopy -->

<owl:Class rdf:about="http://www.semanticweb.org/OntoDomLab-Med#Hysteroscopy">
    <rdfs:subClassOf rdf:resource="http://www.semanticweb.org/OntoDomLab-Med#NonSurgicalProcedure"/>
    <rdfs:subClassOf>
        <owl:Restriction>
            <owl:onProperty rdf:resource="http://www.semanticweb.org/OntoDomLab-Med#ProcedureOf"/>
            <owl:someValuesFrom rdf:resource="http://www.semanticweb.org/OntoDomLab-
            Med#Endometritis"/>
        </owl:Restriction>
    </rdfs:subClassOf>
    <rdfs:subClassOf>
        <owl:Restriction>
            <owl:onProperty rdf:resource="http://www.semanticweb.org/OntoDomLab-
            Med#hasDirectProcedureSite"/>
            <owl:someValuesFrom rdf:resource="http://www.semanticweb.org/OntoDomLab-Med#Uterus"/>
        </owl:Restriction>
    </rdfs:subClassOf>
    <rdfs:subClassOf>
        <owl:Restriction>
            <owl:onProperty rdf:resource="http://www.semanticweb.org/OntoDomLab-Med#hasMethod"/>
            <owl:someValuesFrom rdf:resource="http://www.semanticweb.org/OntoDomLab-Med#Inspection"/>
        </owl:Restriction>
    </rdfs:subClassOf>
    <rdfs:subClassOf>
        <owl:Restriction>
            <owl:onProperty rdf:resource="http://www.semanticweb.org/OntoDomLab-Med#usesDevice"/>
            <owl:someValuesFrom rdf:resource="http://www.semanticweb.org/OntoDomLab-
            Med#Hysteroscope"/>
        </owl:Restriction>
    </rdfs:subClassOf>
</owl:Class>
```

Figure 8 : Définition formelle de <hystéroscopie> lisible par la machine.

6. Lien entre OntDomLab-Med et TEI-Lex-0 appliquée au DLPC

TEI[29] (Text Encoding Initiative), qui est à la base de TEI Lex-0, est une norme *de facto* qui sert à baliser tous types de textes écrits numérisés,

[29] https://tei-c.org/

allant des livres standard aux poèmes, en passant par d'autres documents plus complexes tels que les tableaux, les formules mathématiques, les recettes de cuisine ou encore la notation musicale. Cette norme définit amplement la façon dont les ressources humaines spécifiques doivent être codées, y compris les corpus textuels morphologiquement annotés. Pour la recherche en cours, nous avons suivi les directives TEI qui contiennent un module spécifique pour baliser les dictionnaires (chapitre 9) et autres types de ressources lexicales.

Étant donné que les directives TEI présentent de nombreuses possibilités de balisage, un sous-ensemble plus strict de TEI, TEI Lex-0 (Romary et Tasovac, 2018 ; Bański et al., 2017) est en cours d'élaboration. La préparation de ce format a débuté en 2016, mené par un groupe de travail DARIAH[30] qui est constitué d'experts en ressources lexicales[31]. L'objectif de TEI Lex-0 est d'établir un codage de base et un format cible pour faciliter l'interopérabilité des ressources lexicales à codage hétérogène. TEI Lex-0 ne doit pas être considéré comme *« format that existing TEI dictionaries can be unequivocally transformed to in order to be queried, visualised, or mined uniformly »*.[32]

Les directives TEI et TEI Lex-0 proposent une liste de balises qui servent à marquer l'usage des mots. L'élément `usg` est l'un des constituants de haut niveau dans les directives qui *« contient, dans une entrée de dictionnaire, les informations sur son usage »*[33], dont la fonction est de relier différents niveaux d'information qui constituent la hiérarchie des entrées. Le type de l'information d'usage indiqué par l'attribut `@type` est obligatoire dans TEI-Lex-0.

Par exemple, pour annoter un sens appartenant au domaine de la médecine, une entrée balisée en TEI Lex-0 doit forcément inclure la marque d'usage `usg` :

```
<usg type="domain">Med.</usg>
```

Afin de permettre une annotation correcte pour chaque sens qui se trouve dans le DLPC, il est nécessaire d'établir une liaison entre chacun de ces sens avec une ou plusieurs classes définies dans l'ontologie. Chaque

[30] Digital Research Infrastructure for the Arts and Humanities:www.dariah.eu/activities/working-groups/lexical-resources/
[31] https://dariah-eric.github.io/lexicalresources/pages/TEILex0/TEILex0.html
[32] https://dariah-eric.github.io/lexicalresources/pages/TEILex0/TEILex0.html#index.xml-body.1_div.1
[33] www.tei-c.org/release/doc/tei-p5-doc/en/html/DI.html

classe de l'ontologie est identifiée, sans ambiguïté, par un IRI (Internationalized Resource Identifier). Par conséquent, l'annotation est effectuée en ajoutant des informations supplémentaires à l'élément `usg`. Cette annotation reliera le sens à un groupe de classes. Plus le nombre de classes est grand, plus de résultats pourront être obtenus au moment où l'utilisateur fera sa requête. Cela dit, afficher un résultat avec trop d'informations et un grand nombre de classes peut s'avérer contreproductif. La présentation des résultats avec un nombre élevé de domaines peut ne pas être utile.

C'est pour cette raison que nous prévoyons de maintenir deux types d'annotation : les domaines et sous-domaines qui constituent l'ontologie et les marques de domaines qui permettent l'annotation des unités lexicales. La première servira le raisonneur d'ontologie, la deuxième servira l'utilisateur du dictionnaire.

Il existe deux approches de balisage possibles pour enregistrer la relation de l'élément `usg` avec la classe de l'ontologie : une qui utilise uniquement le format TEI Lex-0 et une autre qui permet l'expansion de TEI Lex-0[34] à savoir la norme W3C XML Linking Language (XLink 1.1).[35]

La première solution a l'avantage d'utiliser uniquement les attributs définis dans TEI Lex-0 ; la deuxième utilise des attributs génériques qui peuvent être appliqués à tous les formats basés sur XML, c'est-à-dire, qui peuvent être utilisés par des outils traitant XML.

En utilisant TEI Lex-0, la balise `usg` possède l'attribut `@corresp` qui peut être utilisé pour pointer *« vers des éléments qui ont une correspondance avec l'élément en question »*[36]. Par conséquent, une relation entre le sens d'« endometrite » [endométrite] (Figure 2) et la classe `disease` [maladie] pourrait être représentée comme suit :

```
<usg type="domain"
  corresp="http://www.semanticweb.org/OntoDomLab-Med#Disease"/>
```

Les éléments de marques des domaines ne sont pas représentés dans le texte du dictionnaire, mais figurent uniquement dans les attributs. Ces éléments ne sont pas représentés dans les articles lexicographiques et donc

[34] XLink n'est pas directement supporté par TEI Lex-0, ce qui veut dire que, pour des propos de validation, il est nécessaire de combiner le schéma TEI Lex-0 et XLink extension.
[35] www.w3.org/TR/xlink11/
[36] www.tei-c.org/release/doc/tei-p5-doc/fr/html/ref-att.global.linking.html

les utilisateurs n'y ont pas accès par défaut, mais ils sont indispensables au raisonneur et ultérieurement pour établir de nouveaux liens entre les entrées.

En revanche, si nous prétendons établir une relation de la classe `Medecine` avec la marque de domaine (*Med.*) qui se trouve dans l'article lexicographique et qui est visible pour l'utilisateur final – une fois que le code XML le transforme en HTML –, cette relation sera annotée comme suit :

```
 <usg type="domain"
   corresp="http://www.semanticweb.org/OntoDomLab-
Med#Medicine">Med.</usg>
```

Comme nous pouvons le constater la marque de domaine (*Med.*) est incluse dans la marque d'usage de haut niveau `usg`.

En utilisant TEI Lex-0 et son expansion en XLink, nous spécifions comment les liens internes et externes entre les différentes composantes structurelles des articles lexicographiques s'établissent.

XLink permet de définir un ensemble d'attributs supplémentaires qui peuvent être utilisés sur n'importe quel élément, nommément `@xlink:href`, qui accepte un IRI du document externe avec lequel on prétend faire la connexion. Ayant comme finalité la définition des marques de domaine pour une entrée, `@xlink:href` comprend l'IRI pour le lier à la classe de l'ontologie. De la même façon que pour `@corresp`, la distinction entre références explicites *(i)* et références implicites *(ii)* pour `@xlink:href` est faite :

```
(i) <usg type="domain"
    xlink:href="http://www.semanticweb.org/OntoDomLab-
Med#Medicine">Med.</usg>
(ii) <usg type="domain"
     xlink:href="http://www.semanticweb.org/OntoDomLab-Med#Disease"/>
```

Le principal avantage de cette approche est que la norme XLink définit l'attribut `@xlink: role`, qui peut être utilisé pour définir la relation entre un certain sens d'une entrée du dictionnaire et des spécifications qui ne sont pas nécessairement une marque de domaine.

```
 <usg type="domain"
   xlink:role="http://www.semanticweb.org/OntoDomLab-
Med#isTreatedBy"
   xlink:href="http://www.semanticweb.org/OntoDomLab-
Med#Gynecology"/>
```

À l'entrée « endometrite » [endométrite] (Figure 2), nous appliquons le TEI-Lex-0 et XLink en associant OntDomLab-Med:

```
<entry xmlns:xlink="http://www.w3.org/1999/xlink"
   type="monolexicalUnit" xml:lang="pt" xml:id="endometrite">
 <form type="lemma">
  <orth>endometrite</orth>
  <pron>ẽdɔmɨtrˈitɨ</pron>
 </form>
 <gramGrp>
  <gram type="pos" norm="NOUN">s.</gram>
  <gram type="gen">f.</gram>
 </gramGrp>
 <!-- etc. -->
 <sense xml:id="endometrite_0" >
  <!-- annotation 1 -->
  <usg type="domain"
    corresp="http://www.semanticweb.org/OntoDomLab-
Med#Medicine">Med.</usg>
  <!-- annotation 2 -->
  <usg type="domain"
    corresp="http://www.semanticweb.org/OntoDomLab-
Med#Inflammation"/>
  <!-- annotation 3 -->
  <usg type="domain"
    xlink:role="http://www.semanticweb.org/OntoDomLab-
Med#isTreatedBy"
    xlink:href="http://www.semanticweb.org/OntoDomLab-
Med#Gynecology"/>
  <def>Inflamação da mucosa uterina.</def>
 </sense>
</entry>
```

Dans cet exemple, il est possible d'observer trois façons différentes d'annoter le même sens en les mettant en rapport avec des domaines ou des classes différentes : les deux premières utilisent uniquement Tei Lex-0, alors que dans le troisième exemple sont appliqués les attributs XLink.

La première annotation fait le lien entre OntoDomLab-Med, par le biais d'une IRI à la classe `Médecine` qui l'associe à la marque de domaine, *Med.*, dans le DPLC.

Dans la deuxième annotation nous ajoutons la classe `Inflammation` qui fait partie de OntDomLab-Med, mais cette information reste

invisible pour l'utilisateur. Cette information peut néanmoins être utilisée par le raisonneur et par le moteur de recherche pour permettre à l'utilisateur de trouver d'autres types d'inflammations, ou de trouver l'entrée « endométrite » à partir de l'unité « inflammation ».

La troisième annotation illustre l'utilisation de `xlink` qui permet de définir une relation spécifique (`isTreatedBy`) avec une classe `Gynécologie` dans l'ontologie. Il convient de noter que l'ontologie comprend différents types de relations (Table 1) qui peuvent être utilisées dans ce contexte pour spécifier clairement la relation entre le sens et toute classe de l'ontologie.

Conclusions

Le développement d'OntoDomLab-Med permet de démontrer la valeur ajoutée apportée par les ontologies appliquées au travail lexicographique. Une telle représentation des connaissances peut contribuer à la création et à l'attribution des marques de domaine plus cohérentes, en plus de permettre une récupération d'informations plus efficace. Étant donné que l'ontologie et les IRI sont par définition indépendants des faits de langue, la ressource ontologique permet de mettre en rapport les concepts avec l'information linguistique multilingue. La valeur ajoutée d'OntoDomLab-Med est donc d'associer l'interopérabilité à la cohérence logique, en connectant l'ontologie et les dictionnaires balisés en TEI-Lex-0. Nous pouvons faire la correspondance entres les marques lexicographiques présentes dans des dictionnaires de langues différentes. Une fois, la méthodologie appliquée à un dictionnaire dans une langue, on créera une méta-étiquette pour opérer comme lien avec les marques correspondantes dans les dictionnaires dans une ou dans d'autres langues.

La méthodologie appliquée et les exemples illustrés dans cet article confirment que l'approche que nous avons adoptée pour la construction d'OntoDomLab-Med constitue une base pertinente pour la modélisation de tout domaine (sciences humaines, sciences sociales, sciences naturelles, etc.), ainsi que pour la modélisation d'autres métadonnées lexicographiques, tels que les registres des langues. Une telle ontologie peut être particulièrement utile pour le traitement des ressources lexicographiques spécialisées, où une gamme plus granulaire de marques

de domaines est nécessaire. Une ontologie de domaines bien définie et un dictionnaire bien balisé permettent de nouvelles approches à l'usage du dictionnaire. Elle permet d'entreprendre des recherches plus restrictives en présentant les sens qui appartiennent exclusivement à un domaine spécifique. Dès lors que les ontologies comportent différents types de relations comme cause/effet, instrument/action, il est possible que certaines références croisées que l'on peut habituellement retrouver au sein des sens des unités lexicales que l'on rencontre dans les dictionnaires soient remplacées ou mises en relief de façon plus accentuée lors de la navigation.

L'organisation de l'information métalinguistique combinée avec un traitement linguistique rigoureux des données permet de repenser le dictionnaire et de le concevoir comme une ressource lexicale réellement adaptée au numérique et non pas comme une reproduction numérisée du dictionnaire papier. Le dictionnaire en tant qu'objet pourra ainsi être envisagé comme une ressource multifonctionnelle répondant de façon plus efficace aux besoins des utilisateurs.

Références bibliographiques

ACADEMIA DAS CIÊNCIAS DE LISBOA, *Dicionário da Língua Portuguesa Contemporânea*, João Malaca Casteleiro (ed.), 2 vols., Lisbonne, Academia das Ciências de Lisboa et Editorial Verbo, 2001.

ARP, R., SMITH, B., SPEAR, D., *Building ontologies with basic formal ontology*, Cambridge, Massachusetts, Londres, Angleterre, MIT Press, 2015.

ATKINS, B. T. S., RUNDELL, M., *The Oxford guide to practical lexicography*, New York, Oxford University Press, 2008.

BAKER, T., BECHHOFER, S., ISAAC, A. H. J. C. A., MILES, A., SCHREIBER, G., SUMMERS, E., SCHREIBER, G., Key choices in the design of Simple Knowledge Organization System (SKOS), *JOURNAL OF WEB SEMANTICS*, 2013, **20**, 35-49.

BÁNSKI, P, BOWERS, J., ERJAVEC, T., TEI Lex-0 guidelines for the encoding of dictionary information on written and spoken forms, *Electronic Lexicography in the 21st Century: Proceedings of eLex 2017 Conference*, 2017, 485-94.

BÉJOINT, H., *The lexicography of English. From origins to present*, Oxford, Oxford University Press, 2010.

BERGENHOLTZ, H., TARP, S., *Manual of specialised lexicography. The preparation of specialised dictionaries*, Amsterdam, John Benjamins Publishing Company, 1995.

BERNERS-LEE, T., HENDLER J., LASSILA, O., The semantic web. A new form of web content that is meaningful to computers will unleash a revolution of new possibilities, *SCIENTIFIC AMERICAN*, 2001, **284**, 1-5.

BORST, W. N., *Construction of engineering ontologies for knowledge sharing and reuse*, University of Twente, 1997.

CARVALHO, S., COSTA, R., ROCHE, C., The role of conceptual relations in the drafting of natural language definitions : an example from the biomedical domain, in KERNERMAN, I., KREK, *Proceedings of the LREC 2018 Workshop Globalex 2018 – Lexicography & WordNets*, 2018, Paris, European Language Resources Association, 10-16.

DERRIERE, S., GRAY, A. J. G., GRAY, N., HESSMAN, F. V., LINDE, T., MARTINEZ, A. P., *Vocabularies in the virtual observatory version 1.19 IVOA recommendation*, 7 octobre 2009.

DOERR, M., Semantic problems of thesaurus mapping, *JOURNAL OF DIGITAL INFORMATION*, 2001, **Vol. 1, No. 8**.

FEDOROVA, I. V., Style and usage labels in learner's dictionaries : ways of optimization, in WILLIAMS, G., VESSIR, S. (ed.), *Proceedings of the 11th Euralex International Congress*, 265-272, 2004, Lorient, France, Université de Bretagne-Sud, Faculté des lettres et des sciences humaines.

GERACI, A., *IEEE standard computer dictionary: compilation of IEEE standard computer glossaries*, IEEE Press, Piscataway, NJ, USA, 1991.

GRUBER, T. R., Toward principles for the design of ontologies used for knowledge sharing, *INTERNATIONAL JOURNAL HUMAN-COMPUTER STUDIES*, 1995, **43(5–6),** 907-928.

GUARINO, N., Formal ontology and information systems, *Proceedings of FOIS '98*, 3-15, Trento, IOS Press, 1998.

GUARINO, N., GIARETTA, P., Ontologies and knowledge bases: towards a terminological clarification, in MARS, N. (ed.), *Towards very large knowledge bases*, 25-32, 1995, Amsterdam, IOS Press.

GUARINO, N., OBERLE, D., STAAB, S., What is an ontology?, in *Handbook on ontologies*, 1–17, Berlin, Heidelberg, Springer Berlin Heidelberg, 2009.

HAUSMANN, F. J., Die Markierung in einem allgemeinen einsprachigen Wörterbuch : eine Übersicht, in HAUSMANN, F. J., REICHMANN, O., WIEGAND, H. E., ZGUSTA, L. (eds.), *Wörterbücher. Ein internationales Handbuch zur Lexikographie*, Berlin/New York, Walter de Gruyter, 1989, 649-657.

ISO/IEC 2382:2015, Information technology – Vocabulary, Recommendation ISO/IEC 2382:2015, Genève, International Organization for Standardization.

KINABLE, D., Reflection on the concept of a scholarly dictionary, *Kernerman Dictionary News*, 2015.

LANDAU, S., Dictionaries, *The art and craft of lexicography*, Cambridge University Press, 2001.

PORPHYRY, Introduction (or Isagoge) to the logical Categories of Aristotle, vol. 2, 609-633. Traduction anglaise par OWEN, O. F., 1853.

PTASZYŃSKI, M. O., Theoretical considerations for the improvement of usage labelling in dictionaries: a combined formal-functional approach, *INTERNATIONAL JOURNAL OF LEXICOGRAPHY*, 2010, **23(4),** 411-442.

REY, A., *Encyclopédies et dictionnaires*, Paris, PUF, Que Sais-je ?, 1982.

ROMARY, L., TASOVAC, T., TEI Lex-0 : a target format for TEI-encoded dictionaries and lexical resources, *Proceedings of the 8th Conference of Japanese Association for Digital Humanities*, 2018, 274-275.

SAKWA, L. N., Problems of usage labelling in English lexicography, *LEXICOS,* 2011, **21**, 305-315.

SALGADO, A., COSTA, R., Marcas temáticas en los diccionarios académicos ibéricos: estudio comparativo, *RILEX – REVISTA SOBRE INVESTIGACIONES LÉXICAS,* 2019, **2 (2),** 37-63.

SALGADO, A., COSTA, R., TASOVAC, T. Improving the consistency of usage labelling in dictionaries with TEI Lex-0, *LEXICOGRAPHY: JOURNAL OF ASIALEX,* 2019a, **6(2),** 133-156.

SALGADO, A., COSTA, R., TASOVAC, T., Mapping domain labels of dictionaries, XIX *EURALEX International Congress: Lexicography for Inclusion*, Alexandroupolis, Grèce, 2020 [à paraître].

SALGADO, A., COSTA, R., TASOVAC, T., SIMÕES, A., TEI Lex-0 in action : improving the encoding of the dictionary of the Academia das Ciências de Lisboa, in KOSEM, I. et al. (eds.), Electronic lexicography in the 21st century. *Proceedings of the eLex 2019 conference* (417-433), 2019a, 1-3 October 2019b, Sintra, Portugal, Brno, Lexical Computing CZ, s.r.o.

STUDER, R., BENJAMINS, R., FENSEL, D., Knowledge engineering: principles and methods, *DATA AND KNOWLEDGE ENGINEERING*, 1998, **25(1-2),** 161-197.

SVENSÉN, B., *A Handbook of lexicography: the theory and practice of dictionary making*, Cambridge, Cambridge University Press, 2009.

THIÉBLIN, E., HAEMMERLÉ, O., HERNANDEZ, N., TROJAHN, C., Survey on complex ontology matching, *SEMANTIC WEB*, 2017, **9(1),** 25-59.

USCHOLD, M., GRUNNINGER, M., Ontologies: principles, methods and applications, *THE KNOWLEDGE ENGINEERING REVIEW*, 1996, **11(2),** 93-136.