KOSonto: An ontology for knowledge organization systems, their constituents, and their referents

Jean Noel Nikiema[1,2,*], Fleur Mougin[3], Vianney Jouhet[4,3] and Stefan Schulz[5,6]

[1]Department of Management, Evaluation and Health Policy, School of Public Health, Université de Montréal, Canada
[2]Centre de recherche en santé publique, Université de Montréal et CIUSSS du Centre-Sud-de-l'Île-de-Montréal, Canada
[3]Univ. Bordeaux, Inserm UMR 1219, Bordeaux Population Health Research Center, team AHeaD, Bordeaux, France
[4]CHU de Bordeaux, Pôle de santé publique, Service d'information médicale, Bordeaux, France
[5]Medical University of Graz, Institute for Medical Informatics, Statistics and Documentation, Graz, Austria
[6]Averbis GmbH, Freiburg, Germany

Abstract

The structure of knowledge organization systems (KOSs) – domain vocabularies, thesauri, terminologies, classification systems, and ontologies – follows different architectural principles and semantic theories. However, many use cases require their integrated use in a given domain. Building a common framework for KOSs is then a prerequisite for any principled account of their use when data annotated by different KOSs should be integrated. We propose an approach rooted in formal ontology, the aim of which is to harmonize the description of the domain itself with the description of the representational artifacts that claim to organize and represent knowledge of this domain. We propose a transparent framework for describing KOSs with a focus on the biomedical domain. Using comprehensive and consistent terminology, we formalize what KOSs represent by introducing KOSonto, an ontology that characterizes representational artifacts on the one hand and describes the relationships to their referents in the domain of application on the other hand. KOSonto uses OWL-DL axioms and is built under BFO and IAO. It accounts for a range of elements that are characteristic of different kinds of KOSs. We illustrate how KOSonto can be used to describe typical biomedical KOSs such as ICD-10, SNOMED CT, and MeSH. Further work will improve the alignment of KOSonto to foundational ontologies and apply this framework to optimize the creation, use, and reuse of mappings between heterogeneous KOSs.

Keywords

Knowledge organization systems, formal ontologies, KOSonto

1. Introduction

For decades, biomedical informatics has focused on artifacts for organizing domain knowledge [1]. Terminologies, controlled vocabularies, dictionaries, thesauri, classifications, nomenclatures, and ontologies – for which we will use the overarching term "**knowledge organization systems**" (KOSs) [2] – vary in scope, granularity, and design principles. Most of them

Proceedings of the International Conference on Biomedical Ontologies 2023, August 28th-September 1st, 2023, Brasilia, Brazil

[*]Corresponding author.

jean.nikiema@umontreal.ca (J. N. Nikiema); fleur.mougin@u-bordeaux.fr (F. Mougin); vianney.jouhet@chu-bordeaux.fr (V. Jouhet); stefan.schulz@medunigraz.at (S. Schulz)
0000-0002-9396-6423 (J. N. Nikiema); 00000-0002-7436-3010 (F. Mougin); 0000-0001-5272-2265 (V. Jouhet); 0000-0001-7222-3287 (S. Schulz)

were created to address specific use cases, often without making their model of meaning explicit. Efforts invested in KOSs have contributed to interoperability and to a better understanding of (classes of) domain entities and the terminological units that refer to them. However, many of the current denominations for KOS are imprecise and misleading. Particularly, the terms “Controlled vocabulary” and “Terminology systems” are misleading because they suggest that these systems mostly describe human language, although they are often applied to KOSs that have a clear focus on the description of domain entities. For example, systems like ICD-10, ChEBI, the Gene Ontology, and the NCBI taxonomy are often referred to as controlled vocabularies although they do not convey any lexical information. Words such as “term”, “concept”, “ontology”, “entity”, “descriptor”, “class”, “property”, and “relation” are used inconsistently, which leads to misunderstandings, particularly in cross-disciplinary cooperations. It also raises the question of why attempts to standardize the entities of a given domain, *viz.* biomedicine and health, are not underpinned by a meta-level standardization of the description of the realm of representation itself, *viz.* its symbols and its relation to the language and the reality of the domain itself.

Only a few *principled analyses* of the domain of representation itself have been carried out. In the 1990s, MIVoc set out to standardize basic semantic notions in medical informatics [3]. Since then, library science, linguistics, philosophy, and the semantic web have fueled knowledge organization activities, without much uptake of MIVoc, which was withdrawn in 2006. Addressing the need for international cooperation, a multilingual scale standardization initiative has been proposed [4], again with only minor generalizations of the usage of this initiative. The result is a confusing mix of approaches, theories, technical terms, and conceptualizations that have been waiting for a thorough cleanup.

It is imperative to carry out this long-awaited cleaning-up process due to the growing importance of diverse KOSs, such as ICD-10, SNOMED CT, MeSH, and MedDRA for biomedical knowledge management. Each of these systems possesses a particular structure, user community, scenarios of use, and representational philosophy. This is particularly pressing where the integration and alignment of KOSs require semantic harmonization between KOSs (and data annotated with them). Indeed, KOSs have always been used in combination. Therefore, despite their structural and semantic heterogeneity, semantic links and correspondences between their elements are sought. There is a long tradition of building bridges between biomedical KOSs. KOS alignment has been an important topic in the semantic web and knowledge graph communities. Numerous heuristics for KOS alignment/mapping/harmonization have been described [5]. In the biomedical domain, the UMLS Metathesaurus [6] has been of great benefit as the longest-running and most enduring effort of KOSs alignment. Additional efforts such as HeTOP [7] and BioPortal [8] are also noteworthy. These resources, driven by practical needs, support some integration of KOSs, without emphasizing their semantic particularities [9]. Only the UMLS Metathesaurus mappings are continually revised manually by domain experts, although there are some automation initiatives [10].

This paper proposes a ontology-based framework for describing KOSs themselves, together with *what they denote.* We propose “KOSonto”, an OWL model under Basic Formal Ontology

(BFO) 2020[1] [11] and using elements of the Information Artifact Ontology (IAO)[2] [12], which is available at https://github.com/JeanNikiema/kosonto. We believe that only after a principled ontological analysis of the constituents of typical KOSs and their representational commitments the value and restrictions of different kinds of KOSs will be sufficiently determined, and the consequences of an alignment between KOSs of different kinds be predicted. A clear and consistent terminology for KOSs themselves should avoid the pitfalls of divergent interpretations of ill-defined words like “ontology”, “concept”, “entity”, “property”, or “knowledge”. We deliberately avoid most of these words (or only use them in a clear context, such as “SNOMED CT concept”).

The paper is structured as follows: Section 2 presents KOSonto based on a KOS content framework; in Section 3, we describe some biomedical KOSs (ICD-10, MeSH, SNOMED CT, and a small HL7 value set) according to KOSonto; and we discuss our main findings in Section 4.

2. KOSonto – The ontology of knowledge organization systems

2.1. Kinds and content of knowledge organization systems

Different elements support the characterization of KOSs as a meaningful ontological category. First, we consider all KOSs **information content entities** (ICEs) according to IAO. ICEs are immaterial but inherent in one or more material bearers [13]. For instance, the content of the KOS (e.g., SNOMED CT) just as of a work of fiction (e.g., Victor Hugo’s “Les Contemplations”), both ICEs, can be stored in many electronic storage systems at the same time. In addition, the constituents of KOSs are also ICEs, e.g., the concept 195967001 – “Asthma (disorder)” or the poem “Vere novo”, respectively. Secondly, KOSs and their constituents are linked to referents (detailed in the following subsection) in a real or fictional world they intend to represent. Finally, they are artifacts and as such created by humans. With these three characteristics, KOSs are **representational artifacts** [14] constituted by a number of **representational units** (**RUs**), which denote particular **referents**. Accordingly, a backbone of a common KOS framework requires a clear characterization of both referents and RUs. KOSonto uses OWL-DL axioms and is classifiable using the HermiT reasoner [15]. The reason for the choice of OWL-DL is popularity, tool support, and human understandability, but also the appropriateness of this language for representing a mostly static domain. KOSonto includes a typology of possible referents of RUs and their ontological foundation, the ontological nature of the RUs themselves, the symbols used in KOSs, and a typology of KOSs.

2.2. The referent as the kind of entity of what is represented in KOSs

A referent is **what** is represented in a KOS, more precisely the thing in the world that is denoted by a RU of a KOS. Everything can be a referent, as the only requirement for being a referent is **to be denoted**. “Referent” is therefore not a meaningful ontological category and not represented in KOSonto. KOSOnto is based on three fundamental categories: particular entities, type entities, and class entities.

[1] https://basic-formal-ontology.org/BFO-2020/
[2] https://github.com/information-artifact-ontology/IAO/

Particular entities[3] are concrete in space and time, have objective existence and ontological significance, and exist independently of human perception or language [16]. We introduce the class *Particular* as the disjunction of *bfo:Continuant* and *bfo:Occurrent*.

Type entities (or **types**) correspond to repeatable, or instantiable (often qualified as abstract) entities. When a type is instantiated by a specific particular, this particular can be referred to as an instance of this type. We divide them further into:

- *Universals as defined by Aristotle*: encompass anything that can be instantiated by particulars. Aristotelian universals are *immanent*, i.e. they exist in their instances, which precludes universals without instances such as unicorns or intergalactic travels.
- *Types by intension*: represent entities of meaning given by means of a formal definition, comparable to the characteristic function in set theory. They do not necessarily extend to things in reality. They allow for defining, e.g., a unicorn as being a pink horse with a single horn, without however claiming its existence. Intensional meanings have classes of particulars as their extensions, including empty classes.
- *Types by extension*: depend on their particular members, without further descriptions. For example, the set {America, Europe, Africa, Asia, Antarctica, Oceania} necessarily and sufficiently corresponds to what is understood by "Continent". However, such representations are not very common in the biomedical domain.
- *Cognitive types*: correspond to mental representations rooted in language and sensory perceptions, regardless of any concrete correlation. A cognitive representation of a particular or conceptual unicorn or the use of the word "centaur" does not mean that unicorns or centaurs exist. We introduce *FictionalType* as a subclass of *CognitiveType* for those things that exist in a fictional world only.

Class entities (or **classes**) are central elements of description logics and are sometimes considered equivalent to types [17]. Their set-theoretic semantics emphasizes the importance of classes of particulars. This is why we grant them a prominent status as siblings of *Particular entities* and *Type entities* and fully define them as the extension of a type that can have only particulars as members:

$$\textit{Class}\ \textbf{equivalentTo}\ (\textbf{extends}\ \textit{some type}\ )\ \text{and}\ (\textbf{has\_member}\ \textit{only Particular}). \qquad (1)$$

In KOSonto, classes are always implemented as classes of OWL particulars, ensuring a coherent and well-defined framework. It manages to sidestep the logical conundrum presented by Russell's paradox [18] and maintains its operational efficacy. Thus, the definition of classes is provided by the set of characteristics of their actual or potential members, the undefined cognitive meaning, or the universal properties inherent in all their members. Whether a class is *currently* or *supposed to be* empty is not a unifying criterion. For example, the classes that extend the types *Unicorn*, *Centaur*, and *Elf* are not identical. Indeed, classes may exist without any particular member having all the characteristics identified to be a member. Classes are *defined* if their necessary and sufficient characteristics allow a particular entity to be recognized as a member of this class; otherwise, the classes are considered *primitive*.

[3] To avoid confusion we highlight that particular entities are modelled as OWL individuals (A-Box elements), but types are also modelled as A-Box entities.

Summing up, types have actual or at least hypothetical instances; the instances of a type are the members of the class it extends to. In OWL, the operator to express class membership is, rather confusingly, **rdf:type**[4]. OWL requires a bi-partition between classes (T-box entities) and individuals (A-box entities). For understanding our model, it is therefore important to be aware of the ontological notion of particular as introduced above and the technical notion of "OWL individuals". Note that all *types* in the ontological sense are therefore modelled in KOSOnto as *OWL individuals*, along with particulars proper. KOSonto introduces the object property **instance_of** as the relation between a particular and a type in the above sense, and **is_a** as the transitive relation between two types (modelled as A-box entities, i.e. OWL individuals). Our example ontology illustrates the parallelism between types and classes as follows. The axiom asserting that a particular that instantiates a type T_1 also instantiates T_2 if T_1 **is_a** T_2 is expressed by the property inclusion "**instance_of ∘ is_a** subPropertyOf **instance_of**".
We have the A-Box entities (OWL individuals) $\textbf{Horse}_{type}$, $\textbf{Vertebrate}_{type}$ and $\textbf{Animal}_{type}$ as members of *AristotelianUniversal*, on which an OWL reasoner (with A-box reasoning) computes the following expected inferences[5]:

Statements on OWL individuals		Reasoner inference	
Bucephalus; Facts: **instance_of** $\textbf{Horse}_{type}$	(2)	**Bucephalus**; Facts: **instance_of** $\textbf{Vertebrate}_{type}$	(5)
$\textbf{Horse}_{type}$; Facts: **is_a** $\textbf{Vertebrate}_{type}$	(3)	**Bucephalus**; Facts: **instance_of** $\textbf{Animal}_{type}$	(6)
$\textbf{Vertebrate}_{type}$; Facts: **is_a** $\textbf{Animal}_{type}$	(4)	$\textbf{Horse}_{type}$; Facts: **is_a** $\textbf{Animal}_{type}$	(7)

We then define OWL classes (T-box entities) based on the hierarchy modeled in the A-box:

Statement		Reasoner inference	
Horse equivalentTo **instance_of** value $\textbf{Horse}_{type}$	(8)	*Horse* subClassOf *Animal*	(11)
Vertebrate equivalentTo **instance_of** value $\textbf{Vertebrate}_{type}$	(9)	*Vertebrate* subClassOf *Animal*	(12)
Animal equivalentTo **instance_of** value $\textbf{Animal}_{type}$	(10)		

Thus, we represent the hierarchical structure of the ontology at the A-box level underneath the *type* hierarchy. In KOSonto, we introduce *FictionalType* subClassOf *CognitiveType*. While *FictionalType* does not require specifying the kind of instances of fictional types, the *FictionalType* class only allows instances of the *InformationContentEntity* (ICE) type. We can therefore distinguish between types of particulars: (i) those that are not ICEs, (ii) those that are ICEs, and (iii) those that are uncommitted: e.g., horses, centaurs (mythical human-horse hybrids), and sumxus (animals of which science disagrees whether only mythical or really existing), or green horses (potential future breeding result). At the A-box level, these entities can nevertheless be linked together, using formal relations such as **is_a** or **instance_of** but also by informal relations such as **is_narrower_than** in addition to the aforementioned relations:

Individual: $\textbf{Sumxus}_{type}$; Facts: **is_a** $\textbf{Vertebrate}_{type}$ (13)

Individual: $\textbf{GreenHorse}_{type}$; Facts: **is_a** $\textbf{Horse}_{type}$ (14)

[4] http://www.w3.org/1999/02/22-rdf-syntax-ns#type

[5] In the equations, classes are depicted in italics, KOSonto relations and A-Box entities are represented in bold, and other parts of the OWL syntax are shown in normal font. Names of OWL individuals that symbolize types have "type" as subscript.

$$\text{Individual: } \textbf{Centaur}_{type}; \qquad \text{Facts: } \textbf{is\_narrower\_than Vertebrate}_{type} \tag{15}$$

$$\text{Individual: } \textbf{Chiron}; \qquad \text{Facts: } \textbf{instance\_of Centaur}_{type} \tag{16}$$

with **Centaur**$_{type}$ being a member of *FictionalType* while **Sumxus**$_{type}$ or **GreenHorse**$_{type}$ could just be members of *CognitiveType* or even *TypeByIntension*. The latter case applies to the scenario where sufficient defining criteria exist, as in the case of the green horse:

$$\textbf{GreenHorse} \text{ equivalentTo } \textbf{instance\_of GreenHorse}_{type} \tag{17}$$

$$\textbf{GreenHorse} \text{ equivalentTo } \textbf{Horse} \text{ and } \textbf{has\_proper\_part} \text{ some } (\textbf{Hair} \text{ and } \textbf{bearer\_of GreenColor}) \tag{18}$$

GreenHorse$_{type}$ is a member of *type*, without commitment to the existence of non-informational instances in reality, other than if it were introduced as a member of *AristotelianUniversal*. It is therefore left open whether the defined class (18) has members. This degressive description is crucial, considering that in certain cases, fictional entities, metaphors, or analogies may be introduced in KOSs to model specific conditions or processes and facilitate the understanding of intricate phenomena. The important concept "Chi" in traditional medicine systems illustrates the need for such a specification. Other examples are mental disorders whose understanding evolves over time in line with new research, clinical insights, and social acceptance (e.g., malleus maleficarum or neurasthenia) or whose existence are denied.

One might argue that types, in the broadest sense, also include relational objects such as predicates or relations, just like mathematical objects in general. A discussion of this is beyond the scope of this paper, in which we refer only to ICEs (cf. subsection 2.3).

2.3. The representational unit as an atomic representation in KOSs

Having presented the range of possible referents for KOS elements – with a digression beyond realism in order to demonstrate the model's flexibility – we now turn to the RUs themselves and their grounding in KOSonto. All RUs are ICEs in the sense of IAO, i.e. generically dependent continuants. What can be considered an atomic form of representation in KOSs varies. We propose a different kind of RUs by using the referent as support of categorization. By answering the question from the perspective of a KOS builder (*Once we have identified the referent, how can we represent it in a KOS?*) and from a KOS user (*For a specific referent, what is its atomic form of representation in a KOS?*), we can identify *three overlapping and inclusive levels* of RUs. Each of these atomic representations can point to a referent. The minimal requirements for RUs are expressed by (19). This corresponds to **level 1**, with a term being a human-readable word or phrase, belonging to a domain-specific vocabulary. If it acts as a preferred label, it must be unique in the KOS. Labels are often artificially constructed (e.g., "Biopsy of head and neck structure") but self-explanatory and unambiguous, regardless of their use in human communication. Although a label can act as a unique identifier, alphanumeric identifiers are more common, apart from the preferred label.

$$\begin{gathered}\textit{RepresentationalUnit}\text{ equivalentTo }\textit{InformationContentEntity}\text{ and}\\ \textbf{proper\_part\_of}\text{ some }\textit{KnowledgeOrganizationSystem}\text{ and}\\ \text{not (}\textbf{has\_proper\_part}\text{ some }\textit{RepresentationalUnit}\text{) and}\\ \text{((}\textbf{has\_proper\_part}\text{ some (}\textit{Literal}\text{ and }\textbf{bearer\_of}\text{ some }\textit{IdentifierRole}\text{) and}\\ \textbf{has\_proper\_part}\text{ some (}\textit{NaturalLanguageTerm}\text{ and }\textbf{bearer\_of}\text{ some }\textit{PreferredLabelRole}\text{)) or}\\ \text{(}\textbf{has\_proper\_part}\text{ some }\textit{OWL\_ClassExpression}\text{))}\end{gathered}\tag{19}$$

In most cases, KOSs offer more than one term per RU (**level 2**), and they play different roles. KOSonto distinguishes *PreferredLabelRole* from *EntryTermRole* with the subclasses *ExactSynonymRole*, *CloseSynonymRole*, *AmbiguousSynonymRole*, *HyponymRole*, *EllipticSynonymRole*. Exact synonyms have the same meaning as the preferred label, and close synonyms have a very similar meaning in the context of the use of this RU. Ambiguous synonyms belong to more than one RU in a KOS, e.g., "lead" for an electric contact or for the chemical element "Pb".

According to the definition provided in [19] for **composite representations**, KOSonto introduces three composite representations: definitions, descriptions, and exemplifications. Definitions provide sufficient and necessary criteria whereas descriptions (also known as elucidations, e.g., in BFO 2020 [11]) provide only necessary criteria. Exemplifications are descriptions by means of concrete examples. RUs with composite representations are introduced in KOSonto as *ExplainedRepresentationalUnit*:

$$\begin{gathered}\textit{ExplainedRepresentationalUnit}\ \text{ equivalentTo }\ \textit{RepresentationalUnit}\text{ and}\\ \textbf{bearer\_of}\text{ some }\textit{CompositeTextualRepresentationRole}\text{ and}\\ \textbf{has\_proper\_part}\text{ some }\textit{Literal}\end{gathered}\tag{20}$$

Another form of representation is an axiomatic representation. RUs may have composite representations as axioms described in a formal language:

$$\begin{gathered}\textit{FormalRepresentationalUnit}\ \text{ equivalentTo }\ \textit{RepresentationalUnit}\text{ and}\\ \textbf{bearer\_of}\text{ some }\textit{DefiniendumRole}\text{ and}\\ \textbf{proper\_part\_of}\text{ some }\textit{LogicalAxiom}\end{gathered}\tag{21}$$

Logical axioms are constituted by logical constructors (symbols and literals) respecting a specific syntax and grammar, e.g., OWL syntax. Logical axioms can be the only atomic representation available for a referent (**level 1**), or provide, with or without textual composite representations, additional information regarding an RU's referent(s) (**level 3**).

We have excluded from our analysis those KOS components that denote relational entities, i.e. in their broadest sense n-ary predicates. Whether they are "first-class" RUs or mere connectors between RUs is controversial. KOSonto includes them as subclasses of *BinaryPredicate* and *TernaryPredicate* and further elaborates on subclasses of these property classes in terms of hierarchy-building predicates, ontological predicates [20], and predicates according to domain / range restrictions in terms of types, particulars, or literals. For example, OWL object properties (*OWL_ObjectProperty*) are ontological relations that hold between particulars, and OWL datatype properties (*OWL_DataTypeProperty*) between particulars and literals. Ternary relations are not supported by OWL [21] and rarely occur in ontologies, with BFO 2020 being a notable exception [11].

3. Application to known biomedical KOSs

In this section, we briefly apply our framework to some reference biomedical KOSs.

ICD-10 is based on a strictly tree-shaped **is_narrower_than** hierarchy[6]. The disjointness of sibling RUs is a fundamental paradigm, expressed by **is_disjoint_with**. Exceptions are RUs named "others", which can be logically described as the complement of the union of their siblings. With existing clinical conditions as their referents, ICD-10 RUs can be described as denoting Aristotelian universals, apart from some examples of epistemic intrusions, such as H40.0 – "Glaucoma suspect". Such ICD-10 RUs could be seen as instantiating *CognitiveType* or alternatively *ICE* (denoting practitioner's knowledge about a patient). Finally, ICD-10 exhibits its own type-to-type relation named "exclusion", which restricts the meaning of a given RU (e.g., *Diabetes mellitus* excludes known cases of *Diabetes in pregnancy*). The knowledge about a patient's condition, as a relevant coding criterion, sheds light on the unclear nature of the referents (diseases, signs, symptoms, or diagnoses) [22]. Finally, ICD-10 exhibits terms described by *EllipticSynonymRole*, i.e. terms that implicitly require their hypernym to be human-understandable. An example is "Lip" as a label for D10.0, which is contextualized by the parent RU D10 – "Benign neoplasm of mouth and pharynx" so that human users intuitively interpret D10.0 as "Benign neoplasm of the lip".

MeSH has an informal tree structure, which can best be represented by the **is_narrower_than** relation – narrower in meaning as defined by SKOS – because the hierarchy encompasses both taxonomic and mereological aspects. Its RUs represent topics in biomedical publications, which are best interpreted as instances of *CognitiveType*. In the context of MeSH trees, RUs have a tree number as an identifier (ID). However, the same term may belong to several trees with different identifiers. Tree-related RUs and tree-independent RUs have to be distinguished. The latter ones (characterized by a separate unique identifier, or UID) can be interpreted as the hypernyms of the former ones. The hierarchical structure of MeSH is therefore more than just the overlay of trees because there is no transitivity between branches of superposed trees via a shared descriptor. As descriptors are ambiguous, there are no unique labels. For example, the descriptor "Nose" has the tree IDs A01.456.505.733, A04.531, and A09.531, as well as the UID: D009666. The coverage of entry terms and free-text definitions is large. Due to the limited granularity of MeSH, many entry terms are not synonyms but hyponyms.

SNOMED CT is a KOS based on OWL-EL and can thus be seen as a hierarchy of classes. All its RUs are *FormalRepresentationalUnit* as they all have axiomatic representations. SNOMED CT allows post-coordination, then also exhibits composite representation as level 1 RUs. SNOMED CT concepts can be seen as extensions of *AristotelianUniversals* or *TypeByIntention*, although in some cases such as 249820005 – "Absence of toe (finding)", a full logical definition, covering the intended meaning of this RU, is not given due to the lack of negation support in OWL-EL. For each RU, most terms are synonyms or near-synonyms with the fully specified names in each language being the preferred label. Textual composite representations are rare. Another particularity is that, in SNOMED CT, everything is named "concept", even those RUs that correspond to binary predicates and which are represented as OWL object properties.

HL7 hl7VS-appointmentReasonCodes, along with many other so-called value sets of the

[6] Most of it may be interpreted as **is_a**

HL7 standard is here presented as an example of a minimalist form of a KOS, consisting only of a *flat list* of RUs, here ROUTINE, WALKIN, CHECKUP, FOLLOWUP, and EMERGENCY. The labels correspond to the ID of each RU. All RUs in this KOS have free-text definitions. The RUs denote Aristotelian universals as they can all be instantiated by a particular appointment.

4. Discussion

Related work. KOSonto is the first, strictly ontology-based, attempt to lay a foundation for a principled ontological account of KOSs, in order to support interoperability and data integration in a domain characterized by the use of numerous KOSs with different structures, semantics, partly overlapping content, and diverging use cases. KOSonto is built under BFO and IAO. In IAO, referents are restricted to particulars, following BFO as an ontology of particulars, although BFO has never clearly committed to the representation of portions of reality beyond particulars. However, not all biomedical KOSs of interest are committed to ontological realism [23] – as claimed for OBO Foundry ontologies – and not all discourses in science and health have only particulars or classes of particulars as referents [24]. KOSonto addresses this by extending BFO beyond the *Continuant* / *Occurrent* bipartition, by introducing the common parent *Particular*, which is then juxtaposed to *Type* and *Class.* The consideration of types as "first-class citizens", besides particulars, can also be found in other upper-level approaches, e.g., Lowe's four-category ontology [25], as well as in the foundational ontologies GFO and UFO [26]. We have also granted this role to *Class* because it is a central element of KOSs and it is an implementation-specific construct in KOSs. Classes facilitate categorization and organization and can be seen as a specific manifestation or implementation of a type, not the type itself. Despite the focus on BFO and IAO in this work – because of the need to represent ICEs – we aim at the compliance of our approach with other ontological frameworks, including the Ontology of General Topology (OGT), and the Descriptive Ontology for Linguistic and Cognitive Engineering (DOLCE). KOSOnto deliberately does not use domain-based frameworks as support, such as the semiotic triad [27] or "Ontoterminology" [28], as it prioritizes practical relevance for KOS harmonization over attempts to achieve a painful reconciliation of the diverging philosophical views.

Main findings. KOSonto and its application to known KOSs highlight how KOSs of the most diverse types can be described: from small, non-hierarchical to mono- and multi-hierarchical systems, from informal to formal ones, and from reality-based to language-centered ones. Besides providing descriptions for the architectural constituents of KOSs, KOSonto particularly accounts for a typing of the different kinds of referents and their relevance for KOSs: types whose instances are ICEs, types whose instances are particulars, types that do not commit to either being hypothetical or real, non-empty classes of particulars or information entities, non-ontological relations such as **is_broader_than**. *Formal KOSs* can be described as KOSs that contain some *FormalRepresentationalUnit* and are expressed in a language based on logic, such as OWL, OBO syntax, or FOL. However, it is important to note that KOSonto has so far centered on referents that focus on *what kind* of entity is referenced, and not on the "targeted referents", i.e. *what particular entity proper* is meant by the use of an RU in a health record [29]. The changing nature and ambiguity raised by the "targeted referent" represent facets associated with the practical use of RUs, which our framework has not specifically addressed for now.

However, by clearly distinguishing different types of referents, it should be compatible with the ideas underlying the cited "Referent Tracking" paradigm. For example, in John Doe's record, the ICD-11 code "2D10.Z" – within an instance of the FHIR resource *Condition* – has a particular cancer in John Doe's thyroid gland as its referent, which is an instance of **Thyroid cancer**$_{type}$ (equally a member of the class *Thyroid cancer*, which is the extension of **Thyroid cancer**$_{type}$). In contrast, in another Jane Doe's record, the same ICD-11 code is present, but in a FHIR *Condition* instance where the slot **verificationStatus** is filled by *Refuted*. Here, the referent is not a particular entity (there is no cancer in Jane Doe's thyroid) but the type **Thyroid cancer**$_{type}$ itself, referred to in a negation statement (an ICE instance). This points to the potential of KOSonto in supporting the work on referent tracking. KOSonto also addresses the complex topic of identifiers, lexical features, and textual composite representations of RUs in KOSs. While many classical ontologies are not concerned with lexical features of RUs beyond unique IDs and human-readable labels (e.g., most Gene Ontology classes do not have synonyms), other KOSs (e.g., MeSH) have their focus on a high coverage of lexical units such as variants, synonyms, and entry terms. When describing KOSs, we consider the lexical properties such as labels, synonyms, and textual composite representations as orthogonal to the ontological aspects. Large parts of SNOMED CT are rich in ontology axioms and also in synonyms and entry terms. Textual composite representations are only present in a small part of SNOMED CT. On the other hand, there are small KOSs, such as the HL7 value set described above, which lack both ontological and lexical richness and are limited to unique labels and textual composite representations. This shows that an easy, mono-axial categorization of KOSs, e.g., in the sense that axiom-rich ontologies are at one end of the spectrum and lexicon-rich informal thesauri at the other, is not satisfactory. Further work will require to align KOSOnto with the descriptions of lexical features of KOSs, such as OntoLex-Lemon [30].

Limitations. We have intentionally left out all aspects of metadata (provenance, version, editorial notes, authors, etc.), as there are already initiatives such as the PROV Ontology[7] or Metadata vocabulary for Ontology Description and publication (MOD) [31]. We deliberately did not elaborate on the quality issues of KOSs either. Particularly, the mismatch between labeling and implicit meaning of KOSs in a particular scenario of use is a known issue [32], particularly due to numerous exclusion rules in classification systems such as ICD-10, so that labels can no longer be interpreted literally. There is also the problem of fuzzy and even misleading labeling of key concepts such as "Clinical finding" or "Qualifier value" in SNOMED CT [33]. Another quality issue is the misuse of formal languages such as OWL to express thesaurus-style content, driven for example by the popularity of the Protégé ontology editor, which seduces users into creating frame-like knowledge models without being aware of the far-reaching consequences of logical inference, cf. [34] for the NCI thesaurus. Similar issues would arise in implementing classification systems like ICD-10 in OWL [35]. A detailed analysis of KOS quality issues is currently not in the scope of KOSonto. On the other hand, our decision to use OWL for the description of KOSonto limited the expressivity of the ontology. This is a pragmatic compromise by the authors who recognize the fact that, despite the reasons mentioned above for using OWL, even OWL is not always well understood and implemented in its full expressivity.

[7] https://www.w3.org/TR/prov-o/

5. Conclusion and future work

By developing KOSonto, we responded to the need for a principled analysis and description of KOSs in the biomedical field. The modeling principles of KOSonto are of important significance, as they lay the foundation for a deeper understanding of how biomedical language and discourse, biomedical entities in reality, and representational artifacts are interconnected. Recognizing that KOSs are an extremely heterogeneous class of knowledge artifacts, built by different communities, for different purposes, on different knowledge organization traditions and using different architectures, it is difficult to foresee a convergent evolution in the near future. Therefore, the need for mapping between different KOSs becomes inevitable, demanding a common framework. The proposed ontology not only characterizes the representational artifacts themselves but also delves into their relationships with a wide range of referents, spanning from real entities to hypothetical and even fictional entities, all of which hold relevance within healthcare and life science discourse. Moving forward, the next crucial phase entails evaluating the suitability of the KOSonto framework for formally describing and facilitating mapping and harmonization activities across diverse KOSs. This evaluation will contribute to the ongoing search for common ground and improve the effectiveness of creating, using, and reusing mappings between heterogeneous KOSs.

6. Acknowledgements

We would like to express our deepest gratitude to the CRESP research center, for their support and provision of accommodations during the development of this article.

References

[1] S. Schulz, J.-M. Rodrigues, et al., Interface terminologies, reference terminologies and aggregation terminologies, Stud Health Technol Inform 245 (2017) 940–944.

[2] A. Isaac, E. Summers, SKOS simple knowledge organization system primer: W3C working group note 18, 2009. URL: https://www.w3.org/TR/skos-primer/.

[3] Medical Informatics Vocabulary (MIVoc). iTeh standards store, 1997. URL: https://standards.iteh.ai/catalog/standards/cen/beb23db9-36ca-4283-aaad-79ff90535f0f/env-12017-1997.

[4] F. Dhombres, J. Charlet, et al., Knowledge representation and management, it's time to integrate, Yearb Med Inform 26 (2017) 148–151.

[5] J. N. Nikiema, V. Jouhet, et al., Integrating cancer diagnosis terminologies based on logical definitions of SNOMED CT concepts, J Biomed Inform 74 (2017) 46–58.

[6] A. T. McCray, S. J. Nelson, The representation of meaning in the UMLS, Methods Inf Med 34 (1995) 193–201.

[7] HeTOP, CISMeF, 1997. URL: https://www.hetop.eu/hetop/.

[8] N. F. Noy, N. H. Shah, et al., BioPortal: ontologies and integrated data resources at the click of a mouse, Nucleic Acids Res 37 (2009) W170–W173.

[9] L. Zheng, Z. He, et al., A review of auditing techniques for the UMLS, J Am Med Inform Assoc 27 (2020) 1625–1638.

[10] G. Bajaj, V. Nguyen, et al., Evaluating biomedical word embeddings for vocabulary alignment at scale in the UMLS Metathesaurus using siamese networks, in: Proc 3rd Workshop on Insights from Negative Results in NLP, 2022, pp. 82–87.

[11] J. N. Otte, J. Beverley, A. Ruttenberg, Basic Formal Ontology, Appl Ontol (2022) 1–27.

[12] B. Smith, W. Ceusters, Aboutness: towards foundations for the information artifact ontology, in: Proc of the 6th Intl Conf on Biomed Ontologies, 2015, pp. 1–5.

[13] E. M. Sanfilippo, Ontologies for information entities, Appl Ontol 16 (2021) 111–135.

[14] B. Smith, W. Kusnierczyk, et al., Towards a reference terminology for ontology research and development in the biomedical domain, in: CEUR Proc, volume 222, 2006, pp. 57–65.

[15] B. Glimm, I. Horrocks, al., HermiT: an OWL 2 reasoner, Journal of automated reasoning 53 (2014) 245–269.

[16] T. Sider, Ontological realism, Metametaphysics (2009) 384–423.

[17] C. M. Fonseca, J. P. A. Almeida, al, Multi-level conceptual modeling: Theory, language and application, Data & Knowledge Engineering 134 (2021) 101894.

[18] A. D. Irvine, H. Deutsch, Russell's paradox (1995).

[19] R. Arp, B. Smith, et al., Building ontologies with BFO, MIT press, 2015.

[20] B. Smith, W. Ceusters, et al., Relations in biomedical ontologies, Gen Biol 6 (2005) 1–15.

[21] R. Hoehndorf, A. Oellrich, et al., Relations as patterns: bridging the gap between OBO and OWL, BMC Bioinformatics 11 (2010) 441.

[22] S. Schulz, J.-M. Rodrigues, et al., What's in a class? Lessons learnt from the ICD–SNOMED CT harmonisation, Stud Health Technol Inform 245 (2014) 1038–1042.

[23] D. Chalmers, Ontological anti-realism, Metametaphysics: New essays on the foundations of ontology (2009) 77–129.

[24] S. Schulz, M. Brochhausen, et al., Higgs bosons, mars missions, and unicorn delusions, in: Proc 2nd Intl Conf on Biomedical Ontologies. CEUR Proc, volume 833, 2011, pp. 183–189.

[25] E. J. Lowe, The four-category ontology, Clarendon Press, 2005.

[26] S. Borgo, A. Galton, et al., Foundational ontologies in action, Appl Ontol 17 (2022) 1–16.

[27] P. T. Raggatt, The dialogical self and thirdness: A semiotic approach to positioning using dialogical triads, Theory & Psychology 20 (2010) 400–419.

[28] C. Roche, Ontoterminology, in: Proc 8th LREC, 2012, pp. 2626–2630.

[29] W. Ceusters, The place of referent tracking in biomedical informatics, in: Terminology, Ontology and their Implementations, Springer, 2022, pp. 39–46.

[30] J. Bosque-Gil, J. Gracia, et al., The OntoLex Lemon lexicography module. Final community group report, 2019.

[31] B. Dutta, A. Toulet, et al., New generation metadata vocabulary for ontology description and publication, in: Metadata and Semantic Research, Springer, 2017, pp. 173–185.

[32] M. Kreuzthaler, M. Brochhausen, et al., Linguistic and ontological challenges of multiple domains contributing to transformed health ecosystems, Front Med 10 (2023) 1073313.

[33] S. Schulz, R. Cornet, et al., Consolidating SNOMED CT's ontological commitment, Appl Ontol 6 (2011) 1–11.

[34] S. Schulz, D. Schober, et al., The pitfalls of thesaurus ontologization–the case of the NCI thesaurus, in: AMIA Annu Symp Proc, 2010, pp. 727–731.

[35] A. Rector, S. Schulz, et al., On beyond Gruber:"Ontologies" in today's biomedical information systems and the limits of OWL, J Biomed Informatics 100 (2019) 100002.