Digital thesauri as semantic treasure troves: a Linguistic Linked Data approach to "A Thesaurus of Old English"

Stolk, S.S.

PART II

Historical Language Thesauri and a Digital Form on the Semantic Web

Chapter 3

3. A digital form for historical language thesauri on the Semantic Web

Computers store knowledge in various digital forms. Each form has its strengths and weaknesses, making some forms better suited for a specific purpose than others. The PDF format, for instance, is highly suitable for storing readable documents, whereas the ZIP format is a better fit for bundling and compressing files. Similarly, the form in which a historical language thesaurus is available can facilitate or hinder its intended use. Offering a thesaurus as a PDF document, for instance, allows users to read it, but hinders automated queries over the content. Similarly, some database technologies with which thesauri can be stored may facilitate querying but be less suited for reuse and expansion of the content. The form that historical language thesauri should be published in on the Web is therefore important to consider and is the main topic of this chapter.

The investigation presented in this chapter was performed in December 2017 and led to the selection of a digital form (constituted by a set of so-called data vocabularies) appropriate for representing historical language thesauri on the Web. The combination and use of these vocabularies has, in a wider community towards standardization of the representation of linguistic and lexicographic resources on the Web, recently been termed Linguistic Linked Data.[1] This name will be used from Chapter 6 onwards to refer to the digital form.

The outline of this chapter is as follows. Section 3.1 reflects on existing publications of historical language thesauri and their shortcomings for both computational efforts and reuse of the knowledge they contain. Best practices for publishing data on the Web are discussed to which, ideally, a data form for historical language thesauri conforms. Subsequently, section 3.2 explores the Semantic Web as a candidate technology for sharing lexicographic material on the Web. Section 3.3 defines a methodology to establish an appropriate Semantic Web form of historical language thesauri. To this end, it lists three criteria for selecting appropriate data vocabularies, sets of terminology in which Semantic Web content can be expressed. Sections 3.4–3.7 then provide analyses of key information components found in thesauri – the topical system, lexical senses, synonymy, and sense attributes – and thereby construct a combination of suitable data vocabularies in which to express such lexicographic resources.

[1] Cimiano et al., *Linguistic Linked Data: Representation, Generation and Applications*. Chapter 7, section 7 of the present thesis discusses recent efforts surrounding Linguistic Linked Data and software developed for exploring resources in this format.

3.1. Historical language thesauri and their current forms

The existing historical language thesauri of Scots and English, discussed in Chapter 1, are offered to readers in paper form or, online, as webpages on the World Wide Web. Both of these forms are geared towards a particular goal: to provide material to users in a readable way. However, for filtering, querying, and other computational processing, these forms are far from suitable. The cause of this disadvantage is, as James McCracken, Principal Editor of the *Oxford English Dictionary*, phrases it, that "human readability always seems bound to conflict with computational parseability at some point".[2] The structure of webpages, for instance, may be machine parseable, but this parseability is meant for visualization of the content in browsers as opposed to other computational processing of the actual knowledge that the webpage documents. In short, both print and webpages are lacking in supporting many of the functional needs for thesauri that were identified in the previous chapter. Webpages are not the only form, though, in which the content of online editions of historical language thesauri exists.

The electronic pages of *TOE*, *HTE*, *HTS*, and *BTH* are created dynamically by combining and arranging bits of information taken from underlying electronic databases.[3] Such practice ensures that it is relatively easy to, for example, change the visualisation of the same knowledge in future updates of the website. A change of logo, a different font, making headings bold rather than underlined — these adjustments are then quickly made. Indeed, it is for such reasons considered good practice to separate presentation from the knowledge behind it.[4] The online editions of historical language thesauri are far from the only lexicographical works that employ a database and generate a presentation on top of it. In fact, as McCracken points out, there is an "increasing concensus" that a separation of knowledge from presentation form "is how good lexicography ought to proceed" — for instance by first compiling a lexical database and afterwards transforming it into a human-readable dictionary or thesaurus (whether in print or electronic).[5]

"Draw[ing] directly on the underlying lexical database" of an electronic lexicographic resource, instead of on its presentation, facilitates computational efforts and reuse of the lexical knowledge in other digital bodies.[6] In other words, the form in which databases of thesauri are published is of importance for their utilization. Unfortunately, the databases behind *TOE*, *HTE*, *HTS*, and *BTH* are not made accessible to the public. As a consequence, executing new queries (such as how many poetic senses are positioned within a given category and all its subcategories) is not possible for users. Additionally, even if these databases were made accessible, there is still the matter of whether their format and access

[2] McCracken, 'The Exploitation of Dictionary Data and Metadata', p. 504.

[3] The digital editions of historical language thesauri maintained by the University of Glasgow, including *TOE4* and *HTE3*, employ MySQL database technology. See the section 'Creation of the *Thesaurus*' of *TOE4*.

[4] See, for instance, the relation between the HTML and CSS standards: 'HTML & CSS'.

[5] McCracken, 'The Exploitation of Dictionary Data and Metadata' p. 504.

[6] Ibid.

protocols facilitate users and enable applications to work with their content.

The digital form in which data is best made available is one of the subjects in guidelines for publishing data on the Web. The FAIR principles, for instance, state a number of requirements to make data findable, accessible, interoperable, and reusable.[7] Similarly, the World Wide Web Consortium (W3C) published their 'Data on the Web Best Practices' (DWBP), which outline how data can be shared to "facilitate interaction between publishers and consumers". Applying these guidelines should net a number of benefits, including reuse and interoperability of the published information. The core advice on digital forms in these guidelines consists of using machine-readable standardized formats for data, reusing common data terminology for the domain, and having persistent identifiers for the dataset and bits of knowledge therein. The databases behind *TOE*, *HTE*, *HTS*, and *BTH* – even if they were to be opened up for public access – do not adhere to the two lastmentioned practices, hampering reuse and interoperability of the information. These databases are based on MySQL technology,[8] which lacks innate features to represent lexical information by reusing common terminology and assigning persistent identifiers in the form of Internationalized Resource Identifiers (IRIs) per database entry (e.g., per lexical sense).[9] IRIs are assigned by the websites, instead, which form stable web addresses (or URLs) for categories – but not for lexical senses or other material – based on information captured in the databases.[10] The question remains, therefore, what data form instead might be better suited to share historical language thesauri on the Web?

3.2. Lexicography and the Semantic Web

In his discussion on the exploitation of dictionary data, James McCracken not only addresses the question of why and how one would want to access such data for further computation, but also what an appropriate digital form would be to achieve such ends. He notes: "It makes sense to adopt a single formalism—if not RDF/OWL, then certainly a model that can be readily converted to and from RDF/OWL".[11] One of the arguments in favour of this formalism is that lexicographical works, McCracken notes, have graph-like characteristics.[12] Lexical senses, for instance, can be seen as nodes of such a graph, connected through such relations as hyponymy and synonymy. By expressing these characteristics in an information graph, the utility of the captured knowledge is enhanced for machine processing.[13] That RDF/OWL is seen as the "standard formalism for information graphs" – and many linguistic

[7] Wilkinson et al., 'The FAIR Guiding Principles for Scientific Data Management and Stewardship'.

[8] See the section 'Creation of the *Thesaurus*' of *TOE4*.

[9] *MySQL 5.7 Reference Manual*.

[10] The exception to this statement is the website of the *HTE* edition incorporated into the *OED*, which coins web addresses that identify lexical senses as well as categories.

[11] McCracken, 'The Exploitation of Dictionary Data and Metadata', p. 512.

[12] Ibid.

[13] Ibid., p. 511.

projects have already successfully adopted it or are in the process of adopting it – merits its recommendation as digital form in which to share lexicographical resources.[14] But what exactly is RDF/OWL and what are the characteristics attributed to this "standard formalism for information graphs"?

RDF and OWL are fundamental standards within the Semantic Web.[15] This Web is, in essence, one of linked data, built on top of a set of data standards and technologies that aim to add "well-defined meaning" to information.[16] These standards and technologies provide an "infrastructure for publishing, storing, retrieving, reusing, integrating, and analyzing data".[17] Its data form, open for anyone to use, is comprised of statements (or triples) that together form a network of information, one in which concepts are identified by IRIs (often in the form of web addresses or URLs).[18] The use of IRIs allows for capturing and identifying data, reusing terminology defined elsewhere, and connecting information found in different digital resources. In effect, this identification mechanism enables thesaurus content to be reused, extended with custom labels and with links to other digital resources. These characteristics of the data form offer intrinsic support for many of the DWBP best practices and facilitate achieving the remaining ones. The underlying data format for Semantic Web information, RDF, is therefore mentioned explicitly in the DWBP documentation as highly suitable for publications on the Web. OWL provides an additional layer of expression that can be used on top of RDF for situations in which highly formal definitions and inferencing mechanisms are required.[19] For more informal levels of expression, other Semantic Web standards, such as SKOS, are an alternative to OWL.[20]

Together, Semantic Web standards offer a well-defined formalism for information graphs. Moreover, they are said to support the functionality desired of historical language thesauri as identified in Chapter 2, since the infrastructure of the Semantic Web allows data to be "retrieved, accessed, reused, and integrated in a meaningful way".[21] Perceived benefits in using this data form are mentioned by Christian Chiarcos et al.[22] One of these benefits is the ability to merge different datasets, or relate different perspectives and conceptualizations of similar data, in order to obtain a combined set of data that is validly formatted.

---

[14] Ibid.

[15] See 'Semantic Web'. For an accessible introduction to the Semantic Web standards and uses, see Allemang and Hendler, *Semantic Web for the Working Ontologist*.

[16] Semantic Web technology was intended to add "well-defined meaning" to information on the Web (Berners-Lee et al., 'The Semantic Web'). Examples of such meaning are relations of hyponomy and of incompatibility. The former can be expressed through the subclassification in the RDFS vocabulary ('RDF Schema 1.1'); the latter through disjointness of classes in the OWL vocabulary ('OWL 2 Web Ontology Language').

[17] Janowicz et al., 'Why the Data Train Needs Semantic Rails', p. 5.

[18] 'RDF 1.1 Concepts and Abstract Syntax'.

[19] 'OWL 2 Web Ontology Language: Document Overview'.

[20] 'SKOS Simple Knowledge Organization System Reference'. This chapter will explore which standards are appropriate for use in a Semantic Web form of historical language thesauri from section 3.3 onwards.

[21] Janowicz et al., 'Why the Data Train Needs Semantic Rails', p. 13.

[22] Chiarcos et al., 'Towards Open Data for Linguistics: Lexical Linked Data'.

Thus, thesauri and sets of data elaborating on them can be queried in unison.[23] A second benefit is an increased level of interoperability. Using standardized terminology in describing linguistic data increases a shared understanding of that data and facilitates their interpretation by software. Moreover, the use of IRIs as identifiers ensures data can be linked without the need for duplication of information from one set into another. The ability to link (or reference) in such a manner is valuable for historical language thesauri, too, since some of these resources are subject to licenses intended for viewing only, stipulating that users are not allowed to copy or download a substantial portion of their content (e.g., *TOE* and *HTE*). By adopting IRIs in published thesauri, their users should be able to explore and extend these resources, engaging with the content offered, without infringing on such licenses. In short, this data form is promising for representing historical language thesauri on the Web. However, since such a thesaurus is yet to be captured in this form, the question remains as to how a Semantic Web form of these lexicographic resources should be obtained.

3.3. Obtaining a Semantic Web form

Considering the existing digital forms of historical language thesauri do not yet include a Semantic Web form, it could be argued that such a form could be modelled in any way one would see fit. However, as DWBP best practice 15 and FAIR principle I2 argue, reuse of common terminology in expressing knowledge increases data interoperability and chances at reuse.[24] Such terminology for expressing the semantics of digital information is found in data vocabularies (which in the Semantic Web and Linked Data communities are sometimes referred to as metamodels, metavocabularies or, simply, vocabularies).[25] DWBP indicates that one can locate appropriate data vocabularies for the Semantic Web through repositories, such as the *Linked Open Vocabularies* (*LOV*) repository.[26] The best practices put forward for Semantic Web data specifically, the 'Best Practices for Publishing Linked Data', also suggest using directories, such as *LOV*, or, alternatively, to look into already published datasets. It is apparent why both documents recommend *LOV*. The repository contains a useful overview of existing Semantic Web data vocabularies, their terminology, and locations of access.[27] Moreover, *LOV* offers functionality to search for individual terms, query over the data vocabularies, and visualise them. At the time of writing, the

[23] This capability facilitates extending original thesaurus content (see requirement R3 in Chapter 2).

[24] See 'Data on the Web Best Practices' and Wilkinson et al., 'The FAIR Guiding Principles for Scientific Data Management and Stewardship'.

[25] DWBP, for instance, employs both *data vocabulary* and *vocabulary*. Articles by Van Assem et al. adopt the term *metamodel* instead. See Van Assem et al., 'A Method for Converting Thesauri to RDF/OWL'; and Van Assem et al., 'A Method to Convert Thesauri to SKOS'.

[26] 'Data on the Web Best Practices'.

[27] *Linked Open Vocabularies*. For an introduction to the website and its usefulness, see the following article: Vandenbussche, 'Linked Open Vocabularies'.

repository lists well over 500 data vocabularies.[28] Knowing where to find existing data vocabularies, one should address the question of which criteria ought to be used in selecting vocabularies for bringing historical language thesauri to the Semantic Web.

Two crucial criteria for selecting a data vocabulary are named by Mark van Assem et al., who have converted many digital resources to Semantic Web forms successfully.[29] Firstly, the chosen terminology should facilitate preserving all knowledge and intended semantics. After all, loss of meaning or misrepresentations of the content may hinder correct interpretations of the content and its reuse. Secondly, the terminology should be a standard (or extend a standard) in order to promote interoperability with other resources similar in nature. In addition to these criteria from Van Assem et al., the present chapter takes a third into account: coherency between terms employed should be well understood. Mixing and matching individual terms from a plethora of data vocabularies, for instance, is likely to net ambiguous or even unknown connections between individual terms. Datasets can be interpreted better when terminology from a few select vocabularies is used instead. In such cases, their cohesion is known and well understood. In short, three criteria are considerd key in selecting appropriate data vocabularies: coverage, standardization, and coherency. Using these three criteria, the next sections will discuss appropriate data vocabularies for each component of historical language thesauri identified in Chapter 1: their topical system, lexical senses, and the relation of synonymy between senses. Sense attributes commonly found in these thesauri (i.e., part of speech, usage features, and language) are covered after these main components.

3.4. Semantic Web form for topical systems

On the Semantic Web, the term "topical system" is not as commonplace as the phenomenon itself. To illustrate, the term *topical* occurs in only four *LOV* vocabularies — not one of which has a clear focus on capturing topical systems.[30] This result begs the question of what terminology is employed on the Semantic Web instead. The answer can be found in one of the meanings of the word *thesaurus*. As Reinhard Hartmann states, one sense of the word is that of a 'terminological database' or 'index'.[31] The *OED* defines this particular sense, first attested in 1957, as follows: "A classified list of terms, esp. key-words, in a particular field, for use in indexing and information retrieval."[32] In essence, this sense of *thesaurus* coincides with what is known as the topical system in historical language thesauri: concepts or labels that are arranged in a hierarchical manner, typically based on the semantics of these concepts, in order to index

[28] This chapter was written in December 2017. The *LOV* repository has grown since and, on 18 April 2022, contains information on 774 vocabularies.

[29] See Van Assem et al., 'A Method for Converting Thesauri to RDF/OWL'; and Van Assem et al., 'A Method to Convert Thesauri to SKOS'.

[30] *LOV*, s.v. 'dbpedia-owl', 'gold', 'lmm1', 'umbel'.

[31] *Encyclopedia of Language & Linguistics*, s.v. 'Thesauruses', by Hartmann.

[32] *OED*, s.v. 'thesaurus, n.', sense 2c.

– or categorize – information in various forms. I will henceforth use the term *indexing thesaurus* to denote this specific sense of *thesaurus*.

Before discussing the results of searching *LOV* for appropriate data vocabularies for indexing thesauri, it should be noted that DWBP already explicitly mentions two data models that are used to capture and exchange such thesauri. These data models, which are considered relatively straightforward since "complex formalisms are most often not needed" for indexing thesauri, are the ISO 25964 data model and W3C's Simple Knowledge Organization System (SKOS) vocabulary.[33] The first-mentioned model is part of the ISO 25964 standard by the International Organization for Standardization. This standard contains guidelines on the development of indexing thesauri and proposes a data model to encourage exchange and interoperability. The current body maintaining the ISO 25964 standard, NISO, recognizes similarities between the international standard (which is divided over two parts) and SKOS.

> ISO 25964-1 essentially advises on the selection and fitting together of concepts, terms and relationships to make a good thesaurus. SKOS addresses the next step, with recommendations on porting the resultant thesauri (or other 'simple Knowledge Organization Systems') to the Web. ISO 25964-2 recommends the sort of mappings that can be established between one KOS and another; SKOS presents a way of expressing these when published to the Web.[34]

The aforementioned sources thus convey that SKOS is considered an appropriate data vocabulary to express indexing thesauri on the Semantic Web. In fact, a data vocabulary has been created specifically to supplement SKOS with terms from ISO 25964 that are not already covered by SKOS, effectively porting the data model of the ISO standard to a Semantic Web context.[35]

Next to SKOS and its ISO 25964 supplement, further data vocabularies exist that treat indexing thesauri. A search for 'thesaurus' in *LOV* yielded a number that merely refer to indexing thesauri rather than expressing them (often recommending their use)[36] or provide a definition for such reference bodies as a whole, lacking terminology to represent the actual content within.[37] Data vocabularies amongst the results that can be used to represent historical language thesaurus content are the Metadata Authority Description Schema (MADS), the Ontopic Ontology (Ontopic), the UNESKOS Vocabulary (UNESKOS), the ISO 25964 SKOS extension (ISO-THES), and, predictably, SKOS. Figure 3.1 depicts these Semantic Web vocabularies, which can be labelled candidate vocabularies for representing a topical system, and the relationship between them.

When considering the candidate vocabularies depicted in Figure 3.1, it is striking that the vast majority of them extends SKOS: these vocabularies complement SKOS with new terminology or specialize terms that already exist in

[33] 'Data on the Web Best Practices'.
[34] 'ISO 25964 Thesaurus Schemas'.
[35] Isaac and De Smedt, *ISO-THES*.
[36] *LOV*, s.v. 'ptop', 'dce', 'dcterms', 'lom', 'edm', 'gndo', 'crm', 'ecrm', 'mtlo'.
[37] These include *LOV*, s.v. 'fabio', 'iol', 'lingvo', 'crm'.

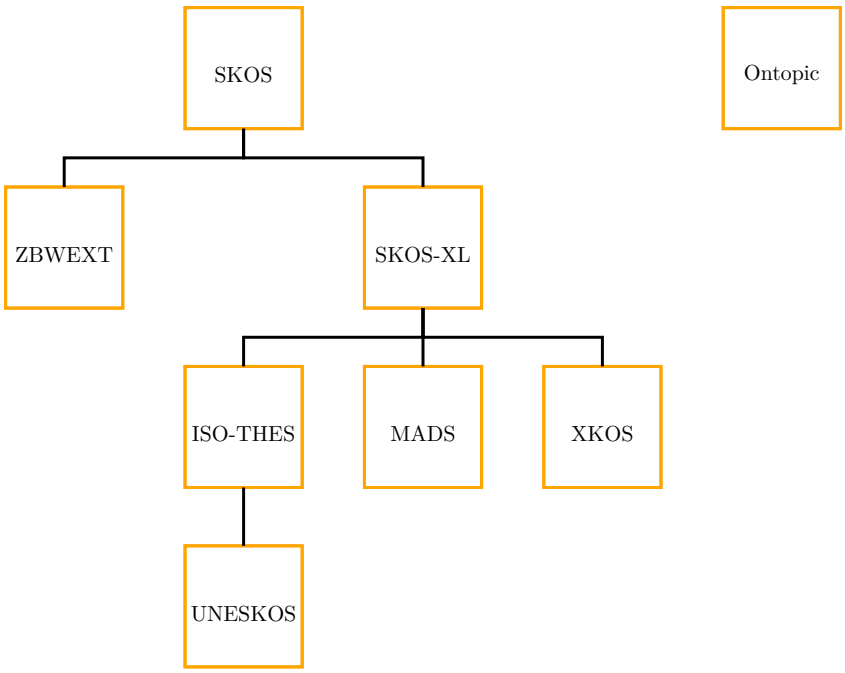

Figure 3.1.: Semantic Web vocabularies that are available for indexing thesauri (depicted as squares). Vocabularies extend higher positioned ones to which they are connected by an upward path.

SKOS. The ISO-THES data vocabulary, for instance, distinguishes three types of hierarchical relations and coins separate terms for these relations, relating each of them to the generic terminology for hierarchical relations available in SKOS.[38] The Ontopic vocabulary is an exception, posited as an alternative to SKOS for indexing thesauri. However, the maintainers of *LOV* remark that, unlike SKOS, Ontopic has seen "no visible use" on the Semantic Web.[39] The prevalence of SKOS and the lack of adoption of Ontopic, which thereby fails to meet the standardization criterion, warrants a closer inspection of SKOS as a vocabulary for expressing topical systems of historical language thesauri.

3.4.1. SKOS for topical systems

The SKOS data vocabulary is meant to express knowledge organization systems. Examples of such systems, according to its specification, are "thesauri, taxonomies, classification schemes and subject heading systems".[40] These systems have a number of traits in common — traits that can be expressed in SKOS in order to bring such knowledge to the Semantic Web environment.

The terminology in SKOS revolves around the notion of concepts. The definition of the term `Concept` is "a unit of thought", a rather general notion

[38] See the properties `iso-thes:broaderGeneric`, `iso-thes:broaderInstantial`, `iso-thes:broaderPartitive`, which are asserted to be subproperties of `skos:broader`.
[39] *LOV*, s.v. 'ontopic'.
[40] 'SKOS Simple Knowledge Organization System Reference'.

that ensures wide support for informal organizing systems.[41] Such systems all express and organize items, and SKOS therefore includes terminology to organize concepts into informal hierarchies (using the relations `broader` and `narrower` between concepts) and to create cross-references between concepts (using the relation `related`). The concepts themselves can be described using labels and notes, and aggregated into collections (`Collection`) and schemes (`ConceptScheme`).

The terminology available in SKOS is expressive enough to capture the majority of the information found in the topical system of historical language thesauri. Each category can be represented as a concept in SKOS, identified by its own IRI, labelled with a name, and placed in an informal hierarchy. The identification of a category, which encodes the position of a category within the topical system, can be expressed using the SKOS `notation` property. These identifications are, in historical language thesauri, sufficient to deduce the order of co-ordinate categories when presenting them. Editorial commentaries, too, can be expressed in SKOS through its system of notes. In short, the coverage of this data vocabulary is extensive and warrants an assessment to determine whether SKOS is sufficiently standardized, too, for representing historical language thesauri.

SKOS was finalized and published in 2009 as a recommended standard by W3C, the consortium that initiated and maintains the technological specifications for the Semantic Web. As such, SKOS is backed by an authoritative body in the Semantic Web community. The quality of the vocabulary is perhaps best illustrated through its use and reuse. As Figure 3.1 shows, SKOS has certainly seen reuse and specialization in other vocabularies, including ISO-THES and XKOS. In fact, *LOV* shows 214 data vocabularies employing SKOS, in one way or another, at the time of writing.[42]

Not just data vocabularies employ terminology from SKOS. A large number of indexing thesauri published on the Semantic Web, too, make use of this data vocabulary. A nonexhaustive list includes EuroVoc (the European Union's multilingual thesaurus), the NASA Thesaurus, the UNESCO Thesaurus, the Getty Vocabularies (including the Art & Architecture Thesaurus and the Getty Thesaurus of Geographic Names), AGROVOC (the United Nation's agricultural thesaurus), the Integrated Public Sector Vocabulary, and the Medical Subject Headings. Van Assem, who has ported several indexing thesauri to the Semantic Web (including the last two mentioned), advocates the use of SKOS in particular over coining new terminology that is completely unrelated to that found in SKOS.[43] Next to having good coverage for historical language thesauri, then, the SKOS data vocabulary is also a standard that has been adopted widely on the Semantic Web.

Although SKOS can express categories, it is not evident as to how category *types* should be captured in this data vocabulary. As mentioned in Chapter 1, the historical language thesauri *TOE*, *HTE*, and *LSM* distinguish such different

[41] Ibid.
[42] *LOV*, s.v. 'skos'.
[43] Van Assem, *Converting and Integrating Vocabularies for the Semantic Web*, pp. 145, 150.

types.[44] *TOE* distinguishes two, *HTE* three, and *LSM* four. Separation of what editors perceived as the macrostructure and the microstructure of the thesaurus, for example, are reflected in the category types.[45] The lack of clarity on capturing these distinctions in SKOS has been recognized by the authors of XKOS. The XKOS data vocabulary extends SKOS in order to express so-called classification levels.

3.4.2. XKOS for classification levels

The XKOS data vocabulary is an extension of SKOS, specifically geared towards meeting the needs of the statistical community for knowledge organization systems.[46] The specific requirements of this community include increased specificity for both hierarchies (distinguishing partonymy from hyponymy) and associations (distinguishing causal, sequential, and temporal relations). Additionally, the statistical community recognises levels in their hierarchical structures. The XKOS specification states that "levels are used as a means to identify concepts within a classification [that are] used to classify instances at the same specificity" — similar to the purpose of category types in historical language thesauri.[47]

Although the purpose of the classification levels in XKOS matches that of category types in historical language thesauri, there is a difference between the two notions. This difference makes XKOS levels unfit for expressing category types. Levels in XKOS "correspond to all those concepts that are same distance from the top of the hierarchy".[48] All top categories belong to the first level, all categories directly subordinate belong to the second level, and so on. This approach is valuable for those thesauri in which categories at a given depth of the taxonomy all share the *type* of category. However, this condition does not always hold for historical language thesauri. As demonstrated in Chapter 1, *TOE*, *HTE*, and *LSM*, distinguish different category types. Some of their categories have subordinates of different category types. In other words, an equal depth in the tree for categories does not imply the same level in the hierarchy of category types. The three thesauri use these types in a manner more flexible than XKOS is able to express: conceptual levels, which reflect the hierarchy of category types, rather than tree levels. This distinction is portrayed in Figure 3.2. Here, although the category "Permission" is four levels deep according to the definition used in XKOS, it is only a single conceptual level deep according to *HTE*. *HTE* categories at this conceptual level in are referred to as, simply, 'Categories'. The category "Disobedience", which is is visualized at the same tree level as "Permission", is nonetheless located in the second conceptual level of *HTE* rather than the first. *HTE* categories at this second conceptual level, which are thesaurus microstructure rather than its macrostructure, are referred to as 'Subcategories'.

[44] See section 1.3.3, 'Identification of categories'.
[45] See section 1.2, 'Main components'.
[46] 'XKOS'.
[47] Ibid.
[48] Ibid.

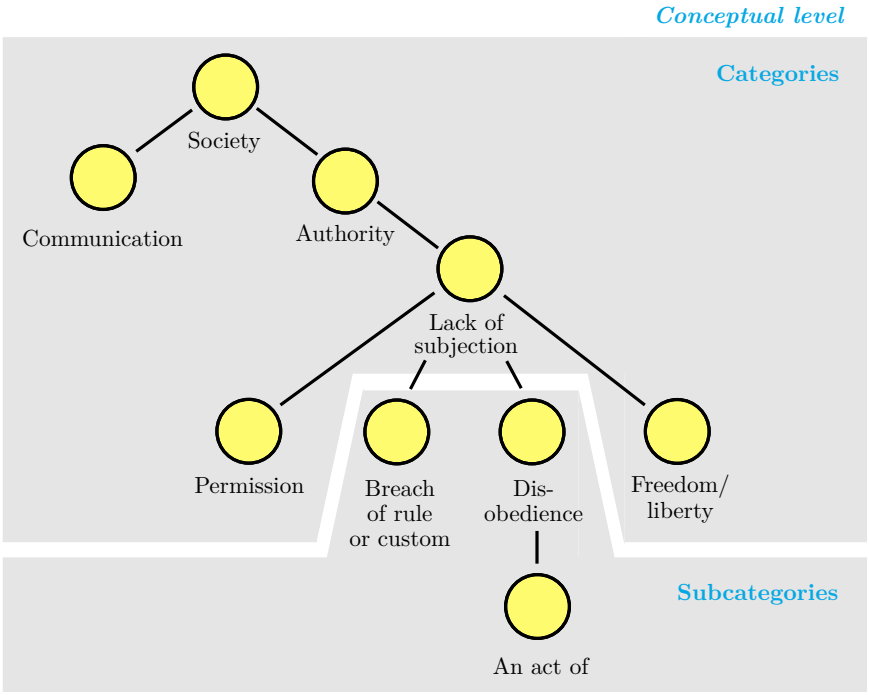

Figure 3.2.: Two conceptual levels in *HTE*.

Although XKOS levels are not appropriate for capturing category types, the manner in which the vocabulary expresses levels offers a valuable insight into how category types can be expressed. Classification levels in XKOS are specializations of concept collections in SKOS. In other words, all concepts that belong to a specific level are asserted to be members of a collection, one that XKOS calls a classification level. A classification level has a certain `depth` (from the top of the hierarchy), typically a name (such as "sections", "subsections", or "divisions"), and possibly a `notationPattern` providing details on the pattern used for the identification of concepts at this level. The concept scheme that contains the concepts may assert the number of levels it contains (using the `numberOfLevels` relation), which these are, and how they are ordered (referring to an ordered list of classification levels through the `levels` relation).

The following can be said on translating the XKOS modelling approach to the context of conceptual levels related to category types found in the historical language thesauri analysed. Firstly, much like classification levels in XKOS, conceptual levels may be posited as subtypes of SKOS collections. Each conceptual level, much like a classification level, typically has a name and may have a notation pattern. The concept schemes that contain conceptual levels would require relations similar to those found in XKOS: a relation to express the number of conceptual levels, and one that lists all available levels in their order of specificity. Such terminology, required for some of the historical language thesauri, are not yet found in the available Semantic Web data vocabularies and

may therefore have to be coined in a new one.[49]

3.5. Semantic Web form for lexical senses

In order to obtain a Semantic Web form for lexical senses, *LOV* has been searched once more for suitable data vocabularies. A search on the keyword 'sense' yields two data vocabularies in the repository that contain terminology for expressing lexical senses: the Lexicon Model for Ontologies (LEMON) and the OLiA Annotation Model for Uby Parts of Speech (UBY). The latter indicates that its term coined for lexical senses (`Sense`) is superseded by that in LEMON (`LexicalSense`), along with ten other terms, such as that for lexicon and lexical entry.[50] For this reason, this section focuses on the LEMON vocabulary and the extent to which it is suitable for capturing senses found in historical language thesauri.

3.5.1. LEMON for lexical senses

The Lexicon Model for Ontologies vocabulary (LEMON) has been designed to capture lexicons and to add their lexicographical knowledge to ontologies in the Semantic Web.[51] The vocabulary has seen a number of updates and was published as a stable W3C vocabulary in May 2016.[52] This finalized version has since been adopted by a number of bodies, including the Global WordNet Association, to represent and link existing lexical resources on the Semantic Web.[53] It is this particular version that the name LEMON will henceforth be used to indicate.

LEMON consists of a number of modules. The core module, called OntoLex, contains terminology that is the most relevant for historical language thesauri. The main terms in this module are those for lexical entries (`LexicalEntry`), lexical senses (`LexicalSense`), forms (`Form`), and lexical concepts (`LexicalConcept`). Reminiscent of dictionaries – and in line with historical language thesauri – a lexical entry has one or more lexical senses and grammatical realizations, or forms. In order to organize lexical entries and senses not alphabetically but onomasiologically, i.e. by their meaning, it is possible to associate them with lexical concepts that can be organized hierarchically.

Lexical concepts in OntoLex are defined as a "mental abstraction, concept or unit of thought that can be lexicalized by a given collection of senses".[54] This definition, not unlike concepts in SKOS, appears highly suitable to express categories from historical language thesauri. In fact, lexical concepts from OntoLex are asserted to be specializations of SKOS concepts. The approach outlined in this chapter to use SKOS for the topical system of thesauri is therefore

[49] See Chapter 5, which introduces *lemon-tree*, a data vocabulary that contains terminology for expressing conceptual levels in thesauri.
[50] *OLiA Annotation Model for Uby Parts of Speech.*
[51] McCrae et al., 'Interchanging Lexical Resources on the Semantic Web'.
[52] 'Lexicon Model for Ontologies'.
[53] 'Global Wordnet Formats'.
[54] 'Lexicon Model for Ontologies'.

strengthened by the specification of LEMON, which makes this connection between the data vocabularies explicit.

As discussed, the LEMON vocabulary covers fundamental terminology for lexical senses in historical language thesauri and their relation to concepts. Next to coverage, the criteria for cohesion and standardization are also met by LEMON. Firstly, the asserted connection with SKOS enables cohesion and increases standardization and interoperability. Secondly, the LEMON data vocabulary has been popular since its creation and has continued to be widely supported. Notable resources that have so far been expressed in LEMON include Princeton's WordNet and Wiktionary,[55] Wikidata,[56] and FrameNet.[57] In short, the framework that LEMON offers is highly suitable for expressing lexical senses from historical language thesauri. Before delving into the matter on which data vocabularies best capture *attributes* of lexical senses, the next section will first discuss how the relation of synonymy, one of the main components found in thesauri, can be expressed.

3.6. Semantic Web form for the relation of synonymy

On the Semantic Web, a small number of available data vocabularies is capable of expressing synonymy as found in thesauri. Searches in *LOV* for 'synonym' and 'synonymy' located a number of candidate data vocabularies for this purpose. The majority of them contains terminology to capture synonymy between elements other than lexical senses.[58] However, as discussed in Chapter 1, the relation of synonymy as found in thesauri is one between senses.[59] Only three vocabularies define their synonymy relations as lexical relations or sense relations: LEMON, GOLD,[60] and LexInfo. The last two vocabularies are positioned as supplementing classifications to LEMON, useful for describing linguistic objects more thoroughly.[61] In order to satisfy the criterion of cohesion, a closer look at LEMON is therefore warranted, since this data vocabulary has already been selected for representing historical language thesauri (i.e., for expressing lexical senses, see section 3.5.1). If LEMON suffices in expressing synonymy for thesauri, its use is therefore preferred over that of GOLD and LexInfo.

In LEMON, the relation of synonymy is asserted not directly between senses but indirectly by grouping them under a lexical concept. Lexical senses that are associated with the same lexical concept through the `isLexicalizedSenseOf` relation are considered to be synonymous. The LEMON specification indicates how its terminology for synonymy corresponds to those used in lexical nets, such as University of Princeton's WordNet, which captures sets of synonyms (referred

[55] McCrae et al., 'Integrating WordNet and Wiktionary with *lemon*'.
[56] Nielsen, 'Lexemes in Wikidata'.
[57] Eckle-Kohler et al., 'lemonUby'.
[58] *LOV*, s.v. 'ru', 'scot', 'dbpedia-owl', 'uniprot'.
[59] Indeed, near-synonymy, which is a form of synonymy between lexical senses, has been called "the staple of thesauruses" (Murphy, 'Meaning Relations in Dictionaries', p. 448).
[60] In full, GOLD stands for General Ontology for Linguistic Description.
[61] 'Lexicon Model for Ontologies'.

to as synsets).[62] The specification thereby demonstrates that synonymy can be captured using solely this data vocabulary, which is shown suitable for capturing this important relation in historical language thesauri on the Semantic Web. With the relation of synonymy covered, this chapter will proceed to discuss which data vocabularies best capture sense attributes.

3.7. Semantic Web form for sense attributes

Most of the historical language thesauri of Scots and English contain information on lexical senses beyond merely the existence of these senses.[63] These thesauri specify such attributes as the part of speech, definition, language, and usage features. Although LEMON suggests definitions are best captured using the `definition` relation from SKOS,[64] choosing an appropriate data vocabulary for expressing the remainder of these attributes on the Semantic Web is not as straightforward as choosing that for the other components has been. Indeed, the LEMON specification states that it neither aims to provide such terminology itself nor wishes to proclaim a single data vocabulary as being the most suitable. Instead, the specification lists a number of efforts that may be useful for describing properties of linguistic objects: GOLD, LexInfo, OLiA, ISOcat (recently superseded by DatCatInfo), and the Clarin Concept Registry.[65] Thus, the specification fails to standardize sense attributes. The reluctance to put forward a single data vocabulary for this purpose is not without reason, as an examination of the expressivity of the existing data vocabularies will show.

3.7.1. Part of speech

The historical language thesaurus *TOE* classifies the verbs it contains as, simply, verbs or as intransitive or transitive verbs. In case of transitive verbs, this thesaurus leaves the number of objects each verb takes unspecified: they could be either monotransitive (and take only a single object) or ditransitive (and take both a direct and indirect object). When reviewing the terminology in the existing data vocabularies, it is not uncommon to find some of the aforementioned verb classifications missing. The GOLD vocabulary, for instance, contains terminology for intransitive verbs, monotransitive verbs, and ditransitive verbs, but lacks terminology for the more general notion of verbs and for transitive verbs of which it is left unspecified whether they are monotransitive or ditransitive. LexInfo, in contrast, has a generic term for verbs, but lacks discrimination in terminology between transitive and intransitive verbs. Some recent initiatives, such as that of the Global WordNet Association, have opted to create further data vocabularies to fulfil their specific needs in capturing parts of speech.[66] There, too, the terminology for parts of speech is incomplete, lacking

[62] Ibid.
[63] As demonstrated in Chapter 1.
[64] 'Lexicon Model for Ontologies'.
[65] Ibid.
[66] See 'Global Wordnet Formats'. The terminology that the association coined for parts of speech has been made available at http://globalwordnet.github.io/schemas/wn.

a distinction between transitive and intransitive verbs (not to mention an absence of such parts of speech as interjections, pronouns, and prepositions). DatCatInfo, employed by the Lexical Markup Framework, is a data category repository that contains definitions for parts of speech, amongst others, and constitutes a rich (albeit non-RDF) alternative to the data vocabularies mentioned above. The repository contains both coarse-grained and fine-grained distinctions for parts of speech, including on transitivity, but lists multiple definitions for a single part of speech, each with its own persistent identifier, and appears to be inconclusive as to which is preferred.[67]

It may well be that there is no one perfect data vocabulary for the parts of speech found across all historical language thesauri. Some of these vocabularies lack specific parts of speech, while others lack (or might even disagree with) the hierarchy between these parts of speech as employed by thesauri. Hans-Jürgen Diller's conclusion, mentioned in Chapter 1 in reference to the topical systems of thesauri, seems relevant for part of speech hierarchies, too: "[t]here is no one right classification; there are only more and less useful ones".[68] Which data vocabulary is best suited to express these particular sense attributes for a historical language thesaurus depends, therefore, on the exact needs of that thesaurus and is best approached on a case-to-case basis.

Regardless of the data vocabulary used, it is evident that the parts of speech together form a system to group lexical items. In that regard, parts of speech are similar to the topical system of thesauri, with two notable differences. Firstly, parts of speech are based on syntactic properties of such items instead of semantic ones. Secondly, they classify rather than categorize content because of the strict and clear criteria that the grouped items have to fulfil. The GOLD and LexInfo vocabularies acknowledge this fact by providing part of speech terminology in the form of a hierarchy of classes.

3.7.2. Usage features

Any attempt at finding a single data vocabulary to capture usage features of senses seems unrealistic. As Chapter 1 has shown, most usage features are indicated through labelling, but the meaning of each label (and how it relates to other labels) is specific to the thesaurus it is found in.[69] A label should therefore always be seen within the defined context of its body. As a result, a shared terminology and definitions for these labels is not likely to be found. LexInfo, for instance, defines a formal register but not an informal one. Moreover, the exact relation between its temporal qualifiers – `archaic`, `obsolete`, and `outdated` – is left unspecified and such relations may very well differ between thesauri. Are

[67] The repository contains four definitions for *verb*, of which two have the status "standardized". Further efforts beyond the Semantic Web towards standardizing parts of speech exist, too. A case in point is the Universal Dependencies framework, which employs the CoNLL-U file format to capture information on sentence tokens, including a universal part of speech tag (e.g., verb, noun) and a language-specific one (see 'CoNLL-U Format', *Universal Dependencies*).

[68] Diller, Review of *HTE1*, p. 322.

[69] Atkins and Rundell, *The Oxford Guide to Practical Lexicography*, pp. 182–6.

outdated items per definition also considered obsolete? Quite as is the case for the part of speech attribute, which vocabulary is best suited to express usage features may depend on the thesaurus in question. If no adequate vocabulary exists, some usage features may even best be represented with terminology specifically coined per thesaurus.

One usage feature that sometimes went beyond mere labelling in the existing historical language thesauri was diachronic marking: stating when a particular sense was in use. This temporal aspect is conveyed in thesauri through named, or even dated, periods in time. One particular data vocabulary, found in *LOV*, appears highly suitable for capturing such temporal aspects: the Time Ontology in OWL (OWL-Time).

OWL-Time for diachronic usage features

The OWL-Time vocabulary contains terminology for expressing temporal aspects and revolves around the notion of temporal entities (`TemporalEntity`).[70] A temporal entity can be either a point in time (`Instant`) or a period with a non-zero duration (`ProperInterval`). In addition to terminology to position such a temporal entity on a given calendar, OWL-Time also contains a number of terms to identify the relative ordering of temporal entities. These terms, based on work by James Allen, include relations to indicate that one interval has taken place before, during, or after another.[71] Thus, temporal aspects can be described quantitively (i.e., with their exact position on a timeline) or qualitatively (i.e., with their relative ordering).

The OWL-Time data vocabulary has been published by the W3C and is, as of June 2017, a candidate recommendation. The combination of LEMON and OWL-Time has already been explored on the Semantic Web. Anas Fahad Khan et al., for example, have employed both data vocabularies to express the temporal extent of lexical domains and semantic shifts.[72] Thus, the Old English period and subperiods thereof could be defined quantitatively (i.e., with exact dates) or qualitatively (i.e., in relation to other periods).[73] Isa Maks et al., too, have opted for this combination of data vocabularies for bringing diachronic lexicons to the Semantic Web.[74]

---

[70] 'Time Ontology in OWL'.

[71] Allen, 'Towards a General Theory of Action and Time'; and Allen and Ferguson, 'Actions and Events in Interval Temporal Logic'.

[72] Khan et al., 'Representing Polysemy and Diachronic Lexico-Semantic Data on the Semantic Web', pp. 42–3.

[73] Ibid., p. 44.

[74] Although the paper in the DHBenelux proceedings still shows a custom data vocabulary used for temporal usage features (Maks et al., "Integrating Diachronous Conceptual Lexicons through Linked Open Data"), the GitHub repository shows that the researchers transitioned to the OWL-Time vocabulary shortly after. https://github.com/cltl/clariah-vocab-conversion/tree/master/rdf-data.

3.7.3. Language

Capturing language, the final sense attribute covered in this chapter, has a more uniform approach than the part of speech and usage features do.[75] LEMON prescribes that the language attribute should be captured using a language tag. Such a tag consists of codes specifically meant to associate a string with a certain language and, possibly, with a specific country or region. The specification requires these language tags to be formed using the language codes based on ISO 639 and, optionally, a hyphen followed by an ISO 3166-1 country code. Thus, this practice requires the "en" tag for English or "en-GB" for English used in Great Britain specifically. Historical languages, too, can be expressed using these codes. Old English, for example, is identified by the language code "ang", Middle English by "enm", and Lowland Scots by "sc". The means to apply these tags to strings are inherent to the RDF format and do not depend on any specific data vocabulary.[76]

3.8. Conclusion

This chapter has discussed what form historical language thesauri should take on the Web. Already available forms of existing historical language thesauri either fall short in terms of the functionality they can offer or hamper reuse and interoperability of their content. A new digital form based on Semantic Web technology may improve the use and reuse of these lexicographic resources. The first step in establishing a Semantic Web form of historical language thesauri has been to select appropriate data vocabularies on the basis of three criteria: coverage, standardization, and coherency. Data vocabularies that meet these criteria have been located for many of the key components found in historical language thesauri of Scots and English. For the topical system, senses, and synonymy, SKOS and LEMON offer terminology that covers most needs. For other components of these thesauri, it has not been as straightforward to find already existing data vocabularies that meet the three criteria. For parts of speech, for instance, there may be no one right data vocabulary to use across all historical language thesauri. Be that as it may, the Semantic Web form constructed in this chapter is in line with prominent guidelines on publishing data on the Web, contributes to data being FAIR, and may well facilitate the functionality over historical language thesauri sought after by researchers — a hypothesis tested in Part III of this dissertation. Before this evaluation, the next two chapters of Part II will continue to discuss aspects of the Semantic Web form

[75] The language attribute is, of course, not solely applicable to lexical senses. Lexical entries, which group such senses, are language-specific, too.

[76] As with use of any registry, there are limitations of which one should be aware: codes registered might not be complete, sufficiently accurate, or sufficiently nuanced. Moreover, one may not be able to ascertain with certainty, especially in contexts of historical languages and the texts that have survived, whether a specific word belongs to a given language. Indeed, a word may even have been misread or misconstructed. Even so, language tags form a good starting point and appear suitable for use with the historical language thesauri listed in Chapter 1.

of historical language thesauri by identifying lacunae in LEMON and introducing complementary terminology for representing thesauri, specifically, in this form.

References

'AGROVOC Linked Data', *AIMS*. http://aims.fao.org/standards/agrovoc/linked-open-data. Accessed: 20 December 2017.

Allemang, D. and J. Hendler, *Semantic Web for the Working Ontologist: Effective Modeling in RDFS and OWL*, 2nd edn (Waltham, 2011).

Allen, J. F., 'Towards a General Theory of Action and Time', *Artificial Intelligence* 23 (1984), 123–54.

Allen, J. F. and G. Ferguson, 'Actions and Events in Interval Temporal Logic', in *Spatial and Temporal Reasoning*, ed. O. Stock (Dordrecht, 1997), pp. 205–45.

Atkins, B. T. S. and M. Rundell, *The Oxford Guide to Practical Lexicography* (Oxford, 2008).

Berners-Lee, T. et al., 'The Semantic Web', *Scientific American Magazine* (May 2001).

Berners-Lee, T., 'Linked Data: Design Issues', *W3C*. https://www.w3.org/DesignIssues/LinkedData.html. Last updated: 18 June 2009.

'Best Practices for Publishing Linked Data', eds. B. Hyland et al., *W3C*. http://www.w3.org/TR/ld-bp/. W3C Working Group Note. Created: 9 January 2014.

Chiarcos, C. et al., 'Towards Open Data for Linguistics: Lexical Linked Data', in *New Trends of Research in Ontologies and Lexical Resources: Ideas, Projects, Systems*, eds. A. Oltramari, P. Vossen, L. Qin, and E. Hovy (Heidelberg, 2013), pp. 7–25.

Cimiano, P. et al., *Linguistic Linked Data: Representation, Generation and Applications* (Cham, 2020). doi: 10.1007/978-3-030-30225-2.

'CoNLL-U Format', *Universal Dependencies*. https://universaldependencies.org/format.html.

'Data on the Web Best Practices', eds. B. F. Lóscio et al., *W3C*. http://www.w3.org/TR/DWBP/. W3C Recommendation. Created: 17 January 2017.

DatCatInfo. https://datcatinfo.net. Accessed: 20 July 2022.

Diller, H., Review of *HTE1*, *Anglia* 128.2 (2010), 319–23.

Eckle-Kohler, J. et al., 'lemonUby: A Large, Interlinked, Syntactically Rich Lexical Resource for Ontologies', *Semantic Web* 6.4 (2015), 371–8.

Encyclopedia of Language & Linguistics, 2nd edn, ed. K. Brown, 14 vols. (Oxford, 2006).

EuroVoc: Multilingual Thesaurus of the European Union. http://eurovoc.europa.eu/. Accessed: 20 December 2017.

General Ontology for Linguistic Description. http://purl.org/linguistics/gold/. Created: 2010.

'Getty Vocabularies', *The Getty Research Institute*. http://www.getty.edu/research/tools/vocabularies/. Accessed: 20 December 2017.

'Global Wordnet Formats', *Global WordNet Association*. http://globalwordnet.

github.io/schemas/. Accessed: 20 December 2017.

Hollink, L. et al., 'Thesaurus Enrichment for Query Expansion in Audiovisual Archives', *Multimedia Tools and Applications* 49.1 (2010): 235–57.

HTE1 = Historical Thesaurus of the Oxford English Dictionary, 2 vols., eds. C. Kay et al. (Oxford, 2009).

'HTML & CSS', *W3C*. http://www.w3.org/standards/webdesign/htmlcss. Accessed: August 15, 2017.

'Integrated Public Sector Vocabulary', *Joinup*. https://joinup.ec.europa.eu/catalogue/asset_release/integrated-public-sector-vocabulary. Accessed: 20 December 2017.

Isaac, A. and J. de Smedt, *ISO-THES*. http://purl.org/iso25964/skos-thes. Created: 17 March 2015.

'ISO 25964 Thesaurus Schemas', *NISO*. http://www.niso.org/schemas/iso25964/. Accessed: 13 June 2017.

Janowicz, K. et al., 'Why the Data Train Needs Semantic Rails', *AI Magazine* 36.1 (2015), 5–14.

Khan, A. F. et al., 'Tools and Instruments for Building and Querying Diachronic Computational Lexica', Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities, Osaka, December 2016, pp. 164–71. https://aclanthology.org/W16-4022.pdf.

Khan, A. F. et al., 'Representing Polysemy and Diachronic Lexico-Semantic Data on the Semantic Web', Proceedings of the 2nd International Workshop on Semantic Web for Scientific Heritage, Heraklion, 30 May 2016, pp. 37–45. https://ceur-ws.org/Vol-1595/paper4.pdf.

'Lexicon Model for Ontologies', eds. P. Cimiano et al., *W3C*. http://www.w3.org/2016/05/ontolex/. Final Community Group Report. Created: 10 May 2016.

LexInfo, 2nd edn. http://www.lexinfo.net/ontology/2.0/lexinfo. Accessed: 20 December 2017.

Linked Open Vocabularies. https://lov.linkeddata.es/. Accessed: 13 June 2017 at http://lov.okfn.org/.

Maks, I. et al., "Integrating Diachronous Conceptual Lexicons through Linked Open Data", Proceedings of DHBenelux 2016, Belval, 8-10 June 2016, pp. 37–45.

McCracken, J., 'The Exploitation of Dictionary Data and Metadata', in *The Oxford Handbook of Lexicography*, ed. P. Durkin (Oxford, 2016), pp. 501–14.

McCrae, J. et al., 'Integrating WordNet and Wiktionary with *lemon*', in *Linked Data in Linguistics: Representing and Connecting Language Data and Language Metadata*, eds. C. Chiarcos et al. (Heidelberg, 2012), pp. 25–34.

McCrae, J. et al., 'Interchanging Lexical Resources on the Semantic Web', *Language Resources and Evaluation* 46.4 (2012), 701–19.

'Medical Subject Headings', *U.S. National Library of Medicine*. https://www.nlm.nih.gov/mesh/. Accessed: 20 December 2017.

Murphy, M. L., 'Meaning Relations in Dictionaries: Hyponymy, Meronymy,

Synonymy, Antonymy, and Contrast', in *The Oxford Handbook of Lexicography*, ed. P. Durkin (Oxford, 2016), pp. 439–56.

MySQL 5.7 Reference Manual. https://dev.mysql.com/doc/refman/5.7/en/. Last updated: 19 December 2017.

'NASA Thesaurus', *NASA*. https://www.sti.nasa.gov/nasa-thesaurus/. Accessed: 20 December 2017.

Navigli, R., 'Word Sense Disambiguation: A Survey', *ACM Computing Surveys* 41.2 (2009), 1–69.

Nielsen, F., 'Lexemes in Wikidata: 2020 Status', Proceedings of the 7th Workshop on Linked Data in Linguistics, Marseille, May 2020, pp. 82-6. https://aclanthology.org/2020.ldl-1.12.pdf.

OLiA Annotation Model for Uby Parts of Speech. http://purl.org/olia/ubyCat.owl. Accessed: 20 December 2017.

'OWL 2 Web Ontology Language: Document Overview', 2nd edn, *W3C*. http://www.w3.org/TR/owl2-overview/. W3C Recommendation. Created: 11 December 2012.

'RDF/OWL Representation of WordNet', eds. M. van Assem et al., *W3C*. http://www.w3.org/TR/wordnet-rdf/. Last updated: 19 June 2006.

'RDF 1.1 Semantics', eds. P. J. Hayes and P. F. Patel-Schneider, *W3C*. https://www.w3.org/TR/rdf11-mt/. W3C Recommendation. Created: 25 February 2014.

'Semantic Web', *W3C*. https://www.w3.org/standards/semanticweb/. Accessed: 13 June 2017.

'SKOS Simple Knowledge Organization System Reference', eds. A. Miles and S. Bechhofer, *W3C*. http://www.w3.org/TR/skos-reference/. W3C Recommendation. Created: 18 August 2009.

'Time Ontology in OWL', eds. S. Cox and C. Little, *W3C*. http://www.w3.org/TR/owl-time/. W3C Candidate Recommendation. Created: 6 June 2017.

'UNESCO Thesaurus', *UNESCO*. http://vocabularies.unesco.org/. Accessed: 20 December 2017.

Vandenbussche, P. et al., 'Linked Open Vocabularies: A Gateway to Reusable Semantic Vocabularies on the Web', *Semantic Web* 8.3 (2017), 437–52.

Van Assem, M., et al., 'A Method for Converting Thesauri to RDF/OWL', *International Semantic Web Conference* (Hiroshima, 2004), 17–31.

Van Assem, M. et al., 'A Method to Convert Thesauri to SKOS', *European Semantic Web Conference* (Budva, 2006), 95–109.

Van Assem, M., *Converting and Integrating Vocabularies for the Semantic Web*. Dissertation DPhil. VU University Amsterdam. 2010. https://research.vu.nl/ws/portalfiles/portal/42185333/complete+dissertation.pdf.

Wilkinson, M. D. et al., 'The FAIR Guiding Principles for Scientific Data Management and Stewardship', *Scientific Data* 3.160018 (2016).

WordNet 3.0 in RDF, eds. M. van Assem and J. van Ossenbruggen. http://semanticweb.cs.vu.nl/lod/wn30/. Last updated: 25 September 2010.

'XKOS: An SKOS Extension for Representing Statistical Classifications', ed. F. Cotton, *DDI Alliance*. http://rdf-vocabulary.ddialliance.org/xkos.html. Unofficial Draft. Created: 18 January 2017.

129

3