Wacana, Journal of the Humanities of Indonesia

Volume 16
Number 2 *From languages to cultures in Indonesia II*

Article 7

7-1-2015

Enterprise vocabulary management; A lexicographic view

Jan Voskuil
guest researcher at McGill University, jan.voskuil@taxonic.com

Follow this and additional works at: https://scholarhub.ui.ac.id/wacana

 Part of the Other Arts and Humanities Commons, and the Other Languages, Societies, and Cultures Commons

Recommended Citation

Voskuil, Jan (2015) "Enterprise vocabulary management; A lexicographic view," *Wacana, Journal of the Humanities of Indonesia*: Vol. 16: No. 2, Article 7.
DOI: 10.17510/wacana.v16i2.384
Available at: https://scholarhub.ui.ac.id/wacana/vol16/iss2/7

This Article is brought to you for free and open access by the Faculty of Humanities at UI Scholars Hub. It has been accepted for inclusion in Wacana, Journal of the Humanities of Indonesia by an authorized editor of UI Scholars Hub.

Wacana Vol. 16 No. 2 (2015): 411–440

Enterprise vocabulary management

A lexicographic view

Jan Voskuil

Abstract

The central theme in this paper is the problem of shifting from natural language descriptions, as in traditional dictionaries and thesauri, to working IT (Information Technology) systems that support people carrying out their administrative tasks. An explicit description of the specific language used in an organization is necessary to guarantee properly working IT systems and a healthy flow of information. Traditionally, there are different ways of capturing such a vocabulary. Different options are considered, arguing that the general form of a thesaurus offers the optimal solution for a broad range of cases. Various requirements for such a thesaurus are examined. A real world example is discussed in some detail. Finally, the paper examines how modern Web technology can help optimizing the creation, management and use of enterprise thesauri. Using these technologies, the enterprise thesaurus can take up new roles in managing the information household of an organization.

Keywords

Vocabulary management, Corporate Language Management, lexicography, thesauri, thesaurus management, IT systems development, SKOS, RDF, Enterprise Data Integration, Data Governance, Data Web, business analysis, Semic Core Vocabularies, semantic relations, taxonomies, knowledge management

1. Introduction

Hardly any human pursuit is thinkable without language. It comes as no surprise, therefore, that the results of linguistic research and practice are applied in so many ways and in so many contexts. This is one of the most fundamental lessons that I learned from Professor Steinhauer: that a scientific view on language can be useful only when it takes an interdisciplinary stance and looks at language as it is actually being used. At the time in the 1980s,

---

Jan Voskuil took his PhD at Leiden University with a study in comparative verb morphology in several Austronesian and Germanic languages. He has been guest researcher at McGill University in Montréal. He started working in the field of IT in the 1990's. He has been involved in several innovative start-ups in the field of Artificial Intelligence, search, and natural language processing. In 2012 he co-founded Taxonic, a Dutch IT consultancy specializing in Linked Data and Vocabulary Management. As a consultant, he is often involved in large projects in public sector organizations. Jan Voskuil can be reached at: jan.voskuil@taxonic.com.

© 2015 Faculty of Humanities, University of Indonesia

this was considered controversial in some circles. After almost 20 years in Information Technology (IT), this lesson is still as valuable to me as it was during my doctoral studies.

In this paper,[1] we explore one such application of linguistics: the construction of thesauri, dictionaries, and glossaries in the context of organizations that are increasingly influenced in their operations by the information technology they use. There is a considerable amount of work taking a technological perspective on structured vocabularies in organizational contexts, and a long tradition in library sciences. This paper takes the intersection of language, language policy, and lexicography as its point of departure.

The central theme in this paper is the problem of shifting from natural language descriptions, as in traditional dictionaries and thesauri, to working IT systems that support people carrying out their administrative tasks. In many organizations, such dictionaries are created by IT people as a side effect of realizing IT systems. In such situations, the quality of the support delivered by the IT system becomes an issue because of terminology problems. More important, however, is: Who decides what counts as the correct definition of a given term? This is a question of language policy at the organizational level. The answer should not be the outcome of ad-hoc problem solving in the operations of IT projects.

Moreover, "data dictionaries" created by the IT supplier do not so much describe the language of the business – that is, of workers in the primary business processes – but rather the inner workings of an IT system. The IT system is supposed to support the business processes, and therefore one can expect it to accurately reflect the language used in these processes. In practice this is an ideal that is difficult to attain, and even in so far as the reflection is faithful in the first place, describing entity and attribute names in an IT system is quite an indirect way of describing the target language.

Therefore, the introduction of large IT systems has had an unexpected side-effect: IT systems determine how the business acts, instead of the other way around.

A concrete example from the world of detention facilities is the following. Inmates who, by developments in society and law, are allowed to stay at their home under a regime of home detention and electronic monitoring, are placed in fictitious cells in a fictitious cell block (often called Z block), because the IT system is designed so that only inmates with a specific cell allocated to them can be registered. It is important to see that there are two anomalies in this

[1] An early version of this paper has been presented at SEMANTiCS 2014 (Voskuil 2014). I am particularly grateful to Thomas Hoppe of Datenlabor Berlin, Claus Blank of Temis Deutschland, and Christine Hörfarter of Geological Survey Austria for discussion. During my time at DJI (2012-2014) I have had ample opportunity to deal with vocabulary management in a practical setting, which is gratefully acknowledged. I am especially indebted to Ruud Binnekamp, Ben Binnendijk, Martien van Bokhorst, Nol van Gemmert, and Lourens Visser. Most of all I am obliged to Bettina Klimek of the Leipzig Institute for Applied Informatics and Jeroen Wiedenhof of the Sinological Institute at the University of Leiden for careful reading of earlier drafts, thorough discussion and good ideas.

real-world example. First, inflexibility of the IT system makes a workaround necessary which in turn leads to an invented, rather contrived concept, namely, that of the "fictitious cell". Second, the real concept of home detention cannot be expressed in this IT system. A thorough analysis of the terms "inmate" and "cell" should have revealed that there is no intrinsic relation between the two. It is easy to point the finger at the IT supplier for assuming that there is, but apparently, the business in its role as authority has not prescribed the correct definitions either.

Anomalies like this abound in any medium sized to large organization. So, the IT system has become a tyrant and a source of inertia instead of a servant and an enabler for change. Creation of a vocabulary in the form of a thesaurus under the aegis of the business can help to shift the balance of power from the IT department back to the business.

In elaborating on this theme, our focus will be on administrative processes. It is in this context that enterprise vocabulary management is especially challenging. Highly specialized work poses its own lexicographic challenges and has enough potential to warrant availability of specialized dictionaries and encyclopaedias. In less specialized, more administrative kinds of work – processing of mortgages, insurances, compliancy, public administration – it is often the case that the vocabulary or jargon that workers are supposed to understand is too complex to grasp at once, but not important enough to warrant professional attention by lexicographers. That is why one can buy specialized jargon dictionaries like a "Dictionary of Legal Terms" or a "Glossary of Cell Biology", but not a "Dictionary of Terms We Use in Organization X".

As a side remark, note that the extent to which specialized jargon dictionaries are published is the result not only of economic but also cultural factors. In Indonesia, where language policy is at the heart of the inception of the nation, a wealth of such dictionaries can be found, often compiled and published by government agencies.[2] An example is the official dictionary of terms in public works (*Kamus istilah bidang pekerjaan umum*), see Figure 1.

The central theme of this paper is how enterprise vocabulary management helps to shift the balance of power from the IT department back to the business. Section 2 describes in more detail the dynamics of how business and IT suppliers (such as the IT department) interact to bring out more clearly the problem of who should be in charge of the corporate vocabulary and the role it should play. Section 3 discusses the different forms in which a vocabulary can be captured and clarifies the concept of thesaurus. Section 4 examines the requirements and good practices for an enterprise thesaurus. Section 5 describes how proven, state of the art Web-technology can be leveraged to realize significant benefits when implementing the enterprise thesaurus. Section 6 provides some reflections and conclusions.

[2] The history of language policy in Indonesia, where more than 500 languages are spoken, is a fascinating story. See for instance Dardjowidjojo (1998) and Steinhauer (2005). Paauw (2009) provides a comprehensive overview of the literature.

Figure 1. *Kamus istilah bidang pekerjaan umum* (Dictionary of terms in public works), published by the Ministry of Public Works in 2008.

2. Business, IT, and language

As noted, the central theme in this paper is the problem of shifting from natural language descriptions, as in traditional dictionaries and thesauri, to working IT systems that support people carrying out their administrative tasks. In this section we consider how business and IT interact in general terms to elucidate the underlying problems.

Organization and business

Organizations exist in all sizes. The primary perspective taken in this paper is the medium-sized organization, with 5,000 to 50,000 employees, with several dozens of IT systems in place, of which at least several have more than a few hundred users. Such organizations are found everywhere: in the public sector, in the financial world, in healthcare, and so on. The ideas in this paper also apply to larger organizations.

Frischmuth et al. (2012) estimate that Volkswagen, with more than half a million employees, has some 5000 IT systems deployed. Such organizations ultimately have the same problems as medium-sized organizations, but at a much larger scale – so that Corporate Language Management (CLM) in that context is a widely recognized phenomenon.[3] CLM is in charge to standardize all corporate terms both for internal and external use, most often taking a multilingual approach. This often results in a variety of unconnected dictionary files for different scopes. However, the need for dictionary creation by the business is recognized and acted upon, whereas in medium-sized organizations this is frequently not the case.

When we talk about "the business", we refer to people involved in or responsible for the execution of the primary business processes in an organization. Organizations deliver certain products or services. To do so, they need to perform certain activities. A business process is a grouping of

[3] See Frischmuth et al. (2012). On CLM, see Fredriksson et al. (2006) and Bursch (2011).

such activities, with a clearly defined start and an end result. Primary business processes directly contribute to delivering the organization's products or services: processing insurances for an insurance company, levying taxes for the tax bureau, and so on. Secondary processes create artifacts that support the primary processes. For example, cleaning the building creates a clean environment, and IT systems development creates an information infrastructure.

Interaction patterns between business and IT

The essence of interaction between the business and suppliers of IT systems is that the business has a need and IT supplies a solution to meet that need. The business is the authority, while IT is the contractor.

In line with this, the interaction proceeds according to a general pattern. The business states the need it has. Business analysts describe the need in more detail. They interview workers involved in or responsible for the business processes and provide a structural description. Next, information analysts create a more detailed description. The next stage is the realization stage, during which the IT solution is realized, either by coding a system or configuring an off-the-shelf product, or a combination of both. When the realization stage is finished and end-user tests are successful, the system can be taken into production. There are many variations on this general pattern. The bottom line in all cases, however, is: the business has a need, and IT supplies a solution.

In this general pattern, it is usually the team of analysts within the IT project that compiles some sort of glossary. This is understandable, as workers in the business are primarily concerned with executing processes, not defining the language used in them. Not every user of a language is a natural born lexicographer, and the same goes for workers in the business. The problem is, however, that business analysts are hardly better suited to the task.

There are two problems conspiring to make this the case: the scope of the business analyst's work within the IT project, and the way IT people including business analysts interpret the notion of lexical meaning, as laid down in a large body of IT literature.

First problem: scope of work

The scope of work in an IT project is clearly delimited. In practice, this means that the context for term definitions comprises limited set of business processes. A holistic overview is necessary to arrive at qualitatively adequate definitions, however.

The root of the problem is that human language is inherently versatile. This makes human communication extremely effective, but in combination with computers this versatility poses challenges. Furnas et al. (1987) point out (in an early research paper on this topic) that people have a strong tendency to use many different words for what from the perspective of IT system development is the same thing. Thus, when different projects address similar

processes, different or even conflicting terms and definitions are unavoidable. The converse is also true of course.

At the Port of Rotterdam, the realization that conflicting definitions were used in different part of the organization prompted an initiative to create a dictionary describing the different uses of terms.4 For instance, the term quay has different meaning for different parts of the organization – waterfront, docking slot, asset. Quay as in "docking slot" can accommodate one ship at a time, whereas quay as in "waterfront" can harbour several. Different usages of the same term may thus imply conflicting definitions. This is in itself not a problem at all: human language is versatile enough to deal with this. It is important, however, to make such definitions explicit when designing IT systems.

An instructive parallel with this is found in law. In many legislations, every law starts with an article listing specific terms and their definition as used in that particular law. Different laws may use conflicting definitions for the same term, or similar definitions for different terms. An adequate dictionary of legal terms must take this as a point of departure, and explain which terms have which definition in which context. This is only possible if a legislation is considered in its entirety. This is what a holistic approach means. IT projects by their very nature do not take a holistic approach. The result is that multiple project-specific glossaries are created with slightly different or outright conflicting definitions for the same terms.

Second problem: lexical meaning construed the IT way

Business analysts working within the context of a specific project are not in a suitable position to create adequate vocabularies. In addition to the lack of overview discussed in the previous section, there is a second reason for this: IT has very different needs when compared to the needs of the business regarding the definition and semantic description of terms.

When creating an IT system or part of it, workers in IT need to create a structural design for storing ("persisting") data and functionality to manipulate these data. The primary focus is therefore in information structure, not on use of language. To make this more concrete, let us consider an example.

A case in point is the set of Semic Core Vocabularies, a European Commission initiative to improve the semantic interoperability of interconnected e-Government systems.[5] In this context, semantic interoperability is defined as the preservation of meaning in the exchange of electronic information. The result is the schema in Figure 2, which represents the conceptual structure of persons and legal entities. To non-IT people this does not quite look like a vocabulary.

[4] I am indebted to L. Visser, at the time Chief Information Officer (CIO) of Port of Rotterdam, personal communication, for bringing this to my attention.

[5] See http://semic.eu.

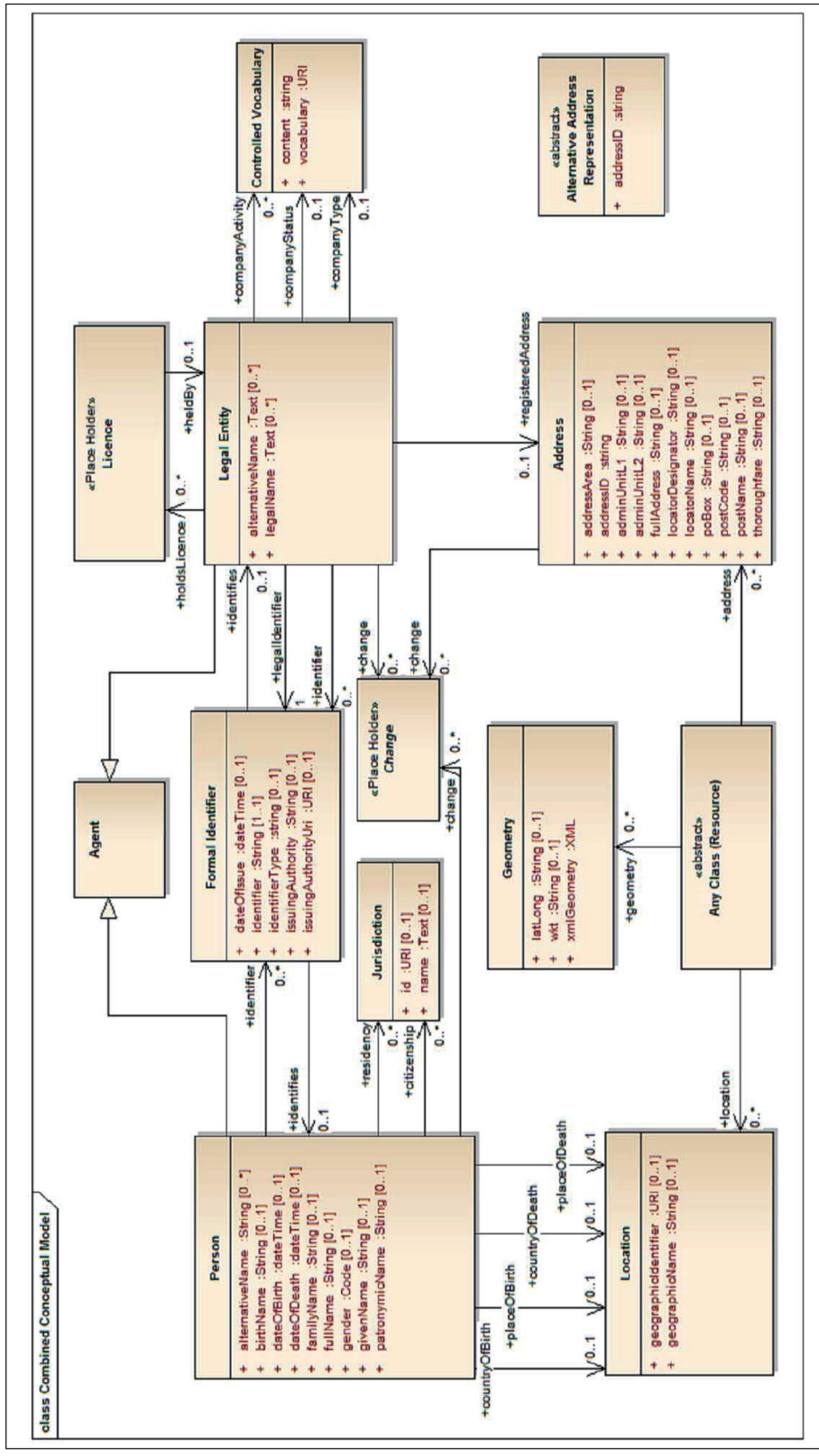


Figure 2. Semic.eu Core Vocabulary.

In the creation of automated IT systems, and of interfaces between them to exchange information, such conceptual schemas are necessary to convey the relations between units of information. For non-IT-people, such schema's contain both too much and too little information. The relations between concepts clutter the picture and often state the obvious: important for design, not so important for general knowledge sharing. At the same time, for business persons it is unclear what the units of information mean. To capture the meaning of these symbols - words, really - one needs definition and clarification in the form of readable text. Take for instance the term "legal entity" in the schema. Different workers may have different ideas on what the term "legal entity" could mean. As long as such differences in interpretation are explicit, they can be dealt with.

Fortunately, the Semic vocabulary has been formalized in an official standard, the Registered Organization Vocabulary, which offers very precise textual definitions with a normative character.[6]

In the day-to-day practice of medium sized organizations, however, IT vocabularies like the one discussed above are not based on such explicit vocabulary definitions. The business does not supply them, and IT initiatives are ill equipped to create these.

IT suppliers supply so-called data-dictionaries as part of the documentation of a system, which do in fact contain textual definitions. However, these are based on the system blueprints and do not aspire to capture business knowledge in any direct way. In particularly bad examples, one finds "definitions" that only state the obvious, like so:

> Person: the person in question. Date of birth: the birth date of the person. Full name: the full name of the person.

For workers in business processes, such data-dictionaries are not helpful.

Conclusion and problem statement

IT suppliers are not well equipped to capture the language of the business. The question is, however, whether they can or even should be expected to do so. Given the role of the business as authority and the IT supplier as contractor, it follows that the business should explicitly capture its language on its own terms. Technical concept schemata and data-dictionaries simply reflect these explicit descriptions, which then give direction to, among other things, IT initiatives. The business leads, the IT supplier follows. Behind this, however, there is the realization that every business process starts with human interaction, and hence, with language. A linguistic norm set by the business is a clear example of organizational language policy, and should be released consciously and deliberately. In the remainder of this paper, we will use the term *enterprise dictionary* to refer to a dictionary created under the aegis of the business to capture the language it uses.

[6] See Archer, Meimaris, and Papantoniou (2013).

Therefore, the problem is: how can one organize matters such that the business creates an enterprise dictionary with clear definitions so that business and IT can be better aligned. In the remainder of this paper, we will develop an answer to this question, starting with a short discussion of the structure of a dictionary.

DICTIONARIES AND THESAURI

Types of dictionaries

Dictionaries have been around for millennia. Among the oldest known are cuneiform tablets with bilingual Sumerian-Akkadian wordlists dated 2300 BCE and an Egyptian thesaurus of medicinal plants found in Thebe dating from 1500 BCE - and there is much more.[7] In the early days, dictionaries were also encyclopaedias and a source of general knowledge. The use of dictionaries exploded in the Renaissance and has ever since been ubiquitous. There is a wide variety of dictionaries: descriptive or normative, alphabetically or thematically ordered, oriented towards spelling or towards correct use, monolingual or bilingual, and within the bilingual category, productive or receptive – to mention just a few distinctions.[8]

For enterprise dictionaries, which aim to capture the language of the business in an organization, the following generic points of departure are particularly relevant:

- The dictionary is normative. It forms an authoritative source of knowledge and defines the correct use of terms within the organization.
- The dictionary is explanatory. It defines terms and describes their use.

In addition, an enterprise dictionary will focus on terms that have a specific meaning within the context of that organization, not on word meanings that can be found in a general dictionary. This is good news, because terms specific to a domain are far easier to define and explain than high frequency words in "ordinary" language. It also means that there is a line to be drawn between words that are incorporated and those that are not. In Section 4 we discuss the ideal size of an enterprise dictionary.

Thematic ordering

A particularly interesting distinction between dictionary forms is the one between alphabetically ordered[9] dictionaries and thematically organized dictionaries, which are also called thesauri. The only valid reason to order entries in a dictionary alphabetically is to make them searchable - or rather, findable. The result, however, is an order that from a lexicographic perspective is completely arbitrary. Throughout the history of dictionary making, there has been a tendency to use a semantics based, thematic ordering instead. Some

[7] On the Egyptian plant thesaurus, see Pavort (2005: 45) and references cited there.

[8] See Atkins and Rundell (2008) for a more thorough discussion.

[9] Note that syllabic, logographic and other writing systems each have conventions for ordering graphemes. The notion of "alphabetic ordering" should be taken in its broadest sense.

of the oldest known dictionaries use thematic ordering and taxonomies, such as the Chinese *Erya* thesaurus from the third century BCE.

Resistance against alphabetic ordering was especially powerful in the nineteenth century, when the Dutch linguist J.W. Muller took up the fight against "the tyranny of the completely unscientific alphabetical ordering". Words such as *spring*, *summer, autumn,* and *winter* are narrower terms of the word *season*. The meaning of any one of these words can only be fully understood in connection with the others. In 1852, P.M. Roget published his *Thesaurus of English words and phrases*. In 1876, Melvil Dewey developed a system for classifying book titles, the famous Dewey Decimal Classification System or DDC for short, which is still in use in libraries today.

Nowadays, thesauri are primarily used in the findability domain for retrieval purposes. Libraries use DDC or a similar system to assign books and other content items to topics. When looking for information on a topic, a user can then retrieve the relevant items. In the days before the computer, libraries used card indexes, nowadays these indexes are computerized. However, the use of thesauri to define the topics and the way these topics relate to each other has not changed. There is a considerable literature on the use of thesauri in the findability domain and library sciences.[10]

A second use for thesauri is to assist users authoring a text in choosing the right word, by presenting synonyms and polysemous alternatives, and antonyms.

Thesauri have become so strongly associated to their application in the findability domain and text authoring, that the important function of explaining the meaning of words has been nudged into the background. Wikipedia goes as far as claiming that "Unlike a dictionary, a thesaurus entry does not give the definition of words",[11] which is in line with much of the literature on taxonomies and thesauri.

This is beside the point, however. Many thesauri have the unambiguous intention to clarify the meaning and use of terms. An example the Art and Architecture Thesaurus Online,[12] which describes thousands of terms like *cathedral* and *impressionist (style)*. Another interesting example is Eurovoc,[13] the multilingual, multidisciplinary thesaurus covering the activities of the European Union (EU). It contains terms and their definitions in 23 EU languages. There are countless other examples of thesauri that offer definitions and explanations of thematically ordered terms. As we will see in Section 5, the widely used SKOS (Simple Knowledge Organization System) standard for thesauri explicitly caters for this function.

In fact, what should be surprising is not that thesauri are used to explain the semantics of terms, but, quite on the contrary, that there are still so many dictionaries around that use alphabetical ordering. Alphabetical ordering is

[10] See for instance Lambe (2007) and Hedden (2011).
[11] See http://en.wikipedia.org/wiki/Thesaurus. Consulted September 2014.
[12] See http://www.getty.edu/research/tools/vocabularies/index.html.
[13] See http://eurovoc.europa.eu.

not necessary to make terms findable, at least not when the dictionary is made available in a digital form. It adds value when, looking at a thesaurus entry, broader terms, narrower terms and related terms are listed as hyperlinks and can be looked up with a mouse click. When one looks up *Impressionist (style)* in the Art and Architecture Thesaurus Online, one not only finds an authoritative description of the term, but also a list of broader terms (among other things, *European styles and periods*) and related terms (*Impressionists (artists)* and *Abstract Impressionist*).

Thesaurus standards

Over the decades, a number of conventions concerning thesaurus construction have gained wide acceptance, and these are formalized in an International Standards Organization (ISO-standard) dating back to 1974.[14] Let us briefly consider these conventions before delving more deeply in the role of thesauri in organizations.

At the conceptual level, the most important features of a thesaurus are that terms are hierarchically ordered. A concept has a preferred label, and one or more alternative labels (AL). The most important semantic relations that are distinguished are Broader Term (BT), Narrower Term (NT), and Related Term (RT). Here are some examples from Eurovoc, the EU thesaurus mentioned previously:

Civil law
- AL: Statutory law
- AL: Ordinary law
- NT: Legal status
- NT: Law of obligations

Civil status
- BT: Legal status
- NT: Civil register
- RT: Marital status

Tax on capital
- NT: Capital transfer tax
- NT: Registration tax
- NT: Wealth Tax

It is interesting to note that the semantic relation NT (and its inverse, BT) is intentionally defined so that it has a highly abstract nature. The standard states: "A hierarchical link between two concepts indicates that one is in some way more general ('broader') than the other ('narrower')".[15] Put differently, it can

[14] See ISO25964 (http://www.niso.org/schemas/iso25964/).

[15] This is actually a quote from the SKOS Reference (Miles and Bechhover 2009, paragraph 8.1). The ISO standard itself defines *narrower term* even more cryptically as the

mean many things in practice.

In some cases, it refers to the "is a type of" relation, so that *butterfly*, *beetle*, and *ant* are narrower terms of *insect*. In other cases, it refers to the relation "is a subtheme of", so that *penal institution* is a narrower term of *criminal law* – of course, one cannot meaningfully say that a penal institution is a type of criminal law!

A third example of how the NT-relation can be used is to refer to the relation between a feature and its value, so that *male* and *female* (the values) can be said to be narrower terms of *sex* (the feature).

The abstractness of the NT-relation – which one may call generalized classification – is actually a very powerful and beneficial property of the ISO-standard. It caters for a kind of "just enough semantics", which is the main ingredient of the polysemy which makes human language so versatile. It adds enough additional structure over an alphabetically organized vocabulary to significantly improve understandability and usability. At the same time, the structure is simple enough to grasp intuitively. One does not have to learn an overkill of highly precise relations – this is exactly what makes the schema in Figure 2 so difficult to understand for business persons.

The semantics of the relation RT, which really means "see also", is similarly kept vague. Precisely because it is so unspecific, it is highly useful in a general purpose vocabulary.

4. The enterprise thesaurus

This section discusses the benefits that an enterprise thesaurus can bring, the conditions that must be met for a thesaurus to be effective, and some good practices regarding size, governance, and skills.[16]

Benefits for the organization

Organizations that take up the task of developing an enterprise dictionary or, better still, an enterprise thesaurus, realize a number of concrete benefits in the following areas:

- Knowledge management. Workers in the primary processes are supposed to know the vocabulary and use it correctly. For new workers, or for experienced employees taking up tasks in new or innovated business processes, the thesaurus serves as an easy to use knowledge base and a trusted source of information. Different meanings of the same term are made explicit and cease to cause confusion.
- Quality of information. An authoritative thesaurus fosters good governance because workers everywhere in the organization speak the same language. Management information and business intelligence are thus more accurate and more reliable.

"preferred term representing a concept that is narrower than the one in question".

[16] I am indebted to Mr. L. Visser, CIO at the Dutch Custodial Institutions Agency, for discussing most of the topics examined in this chapter and offering practical reflections.

- Connecting information. Transparency to the public, to shareholders and supervisory bodies requires a clearly defined vocabulary as a basis.[17] Exchange of information using automated interfaces across organizational boundaries requires semantic standards. An organization needs to have explicit knowledge of its own vocabulary, so that it can connect its own terms and shared terms defined in the standard. Without an organizational vocabulary in place, such mapping is impossible.
- Alignment of business processes and IT systems. The thesaurus gives direction to initiatives in the IT department by providing a solid foundation for designing IT solutions. The organizational thesaurus is the single source of truth when it comes to finding the correct terms and definitions. This speeds up processes in the IT department. Analysts do not have to describe the meaning of central concepts over and again, let alone deal with conflicting views among different sources. In addition, the thesaurus results in qualitative benefits in the form of IT systems that fit the business processes more closely. It prevents a proliferation of different terms naming the same things in different IT systems.

A real world example

Let us consider a real-world example. DJI (Dienst Justitiële Inrichtingen) is the Dutch Custodial Institutions Agency. It manages about 50 detention facilities with 12,000 employees and some 70,000 individual inmates on a yearly basis, the majority of which stays only for several days or weeks. For the purpose of transparency, the organization publishes a number of resources every year describing various aspects of the work it does. A case in point is "Gevangeniswezen in Getal 2008-2013" (Prison Services in Numbers 2008-2013), which can be downloaded for free from the organization's Website.[18] The audience it targets is workers in the domain, policy makers, researchers, and other interested parties. Some 10 pages of the 77 pages that constitute the report compile a glossary of terms.

The definitions in this glossary clearly have the function to provide the outside world with knowledge so that it can understand the information presented in the report. From the above description it is also clear that part of the target audience is workers in the primary processes. They are supposed not only to know what the terms mean but also to actively use them in their daily work. The quality of the information in the report directly depends on the accuracy of the definitions in the glossary. Finally, the glossary plays an obvious role in the alignment of business and IT. The report's colophon states that the report including the glossary is authored by a department involved in the primary processes at the strategic level.[19] Ownership of the glossary

[17] See Santosuosso and Malerba (2014), who argue for distinguishing organizational, legal, and cultural interoperability alongside systems interoperability, for extensive discussion of this point.

[18] See DJI (2014).

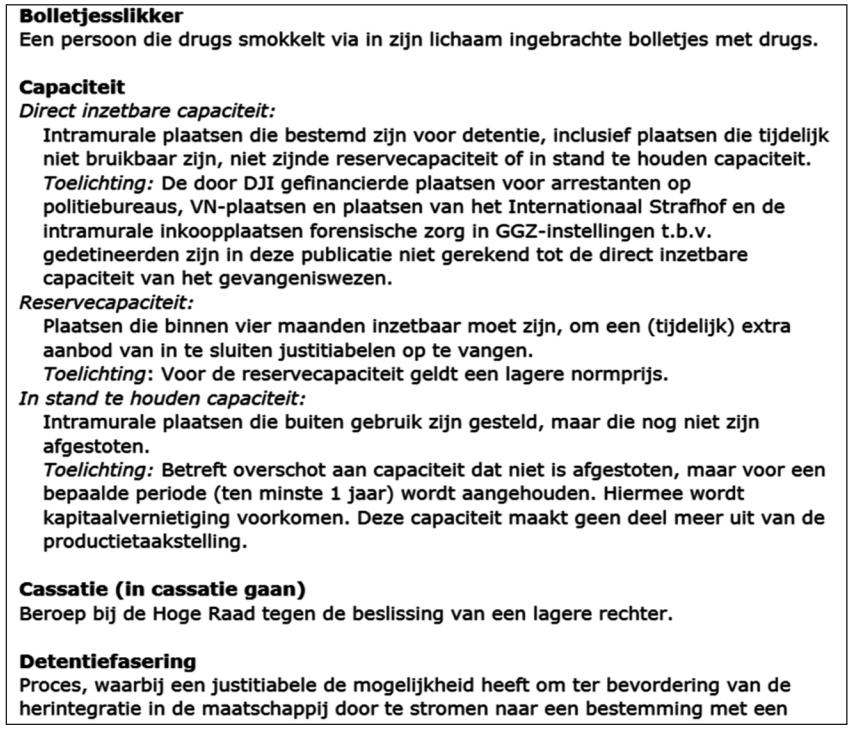
[19] To be precise: the Department of Analysis, Strategy and Knowledge of the Directorate

lies with the business, therefore. It can be expected that the terms and their definitions are leading in IT initiatives.

The importance of a thematic ordering

Interestingly, when considered in more detail, the glossary actually appears to have the structure of a thesaurus. Some terms are described at locations that correspond to a thematic ordering rather than an alphabetical ordering. For instance, the terms *Direct inzetbare capaciteit* (directly deployable capacity), *Reservecapaciteit* (backup capacity) and *In stand te houden capaciteit* (suspended capacity) are printed underneath the broader term *Capaciteit* (capacity). To obtain this result, the glossary uses two typographical tricks:

- The top level terms are bold faced. Narrower terms are printed in italics, and the defining text is indented. See Figure 3.
- The narrower terms are also printed in bold face at the position where they belong alphabetically. For instance, *Reservecapaciteit* is also printed after *Regimesgebonden verlof*. However, at that position the glossary only refers to the broader term: "see *Capaciteit*". Of course, this is not intended to mean that *reservecapaciteit* is a synonym of *Capaciteit* (which obviously it is not) - it only serves the purpose of making the term and its definition findable by alphabetic order.



Bolletjesslikker
Een persoon die drugs smokkelt via in zijn lichaam ingebrachte bolletjes met drugs.

Capaciteit
Direct inzetbare capaciteit:
Intramurale plaatsen die bestemd zijn voor detentie, inclusief plaatsen die tijdelijk niet bruikbaar zijn, niet zijnde reservecapaciteit of in stand te houden capaciteit.
Toelichting: De door DJI gefinancierde plaatsen voor arrestanten op politiebureaus, VN-plaatsen en plaatsen van het Internationaal Strafhof en de intramurale inkoopplaatsen forensische zorg in GGZ-instellingen t.b.v. gedetineerden zijn in deze publicatie niet gerekend tot de direct inzetbare capaciteit van het gevangeniswezen.
Reservecapaciteit:
Plaatsen die binnen vier maanden inzetbaar moet zijn, om een (tijdelijk) extra aanbod van in te sluiten justitiabelen op te vangen.
Toelichting: Voor de reservecapaciteit geldt een lagere normprijs.
In stand te houden capaciteit:
Intramurale plaatsen die buiten gebruik zijn gesteld, maar die nog niet zijn afgestoten.
Toelichting: Betreft overschot aan capaciteit dat niet is afgestoten, maar voor een bepaalde periode (ten minste 1 jaar) wordt aangehouden. Hiermee wordt kapitaalvernietiging voorkomen. Deze capaciteit maakt geen deel meer uit van de productietaakstelling.

Cassatie (in cassatie gaan)
Beroep bij de Hoge Raad tegen de beslissing van een lagere rechter.

Detentiefasering
Proces, waarbij een justitiabele de mogelijkheid heeft om ter bevordering van de herintegratie in de maatschappij door te stromen naar een bestemming met een

Figure 3. Excerpt of the glossary in "Gevangeniswezen in getal 2008-2014".

of Policy Support. One of this department's functions is to provide business intelligence, of which the yearly trend analysis is a result.

The reason for taking the trouble of combining the structure of a thesaurus with alphabetical ordering is obvious: to understand the concept of capacity as used within the organization, one needs to understand its narrower terms in relation to each other. Using the conventions for structuring a thesaurus discussed at the end of the previous section, the structure can be conceptualized as follows:

capacity
NT: Directly deployable capacity
NT: Backup capacity
NT: Suspended capacity

Thus, the thematic order of the vocabulary adds the additional practical benefit of giving contextual information.

Requirements for an enterprise thesaurus

The glossary of terms discussed in the previous paragraph is actually part of DJI's enterprise dictionary. Though it is not published, it does have an official, normative status as it is approved by the organization's Senior Executive Board. Nevertheless, the situation leaves room for improvement:[20]

- Coverage of the dictionary is limited: it only describes slightly more than a 100 terms.
- As a consequence, there are more glossaries around, often with a project specific scope.
- The dictionary is available as a PDF-file only and is distributed by mail. This causes limited uptake within the organization.
- There is no regular maintenance process in place, so that new terminology introduced by projects has to be dealt with in an ad-hoc fashion, if at all.

This suboptimal situation is representative for most medium-sized organizations. For an enterprise thesaurus to fully realize its potential benefits, the following criteria must be met:

- The thesaurus is available for all employees on the intranet, or preferably even to stakeholders outside the organization on the organization's public Website. This guarantees ease of access.
- The thesaurus has enough coverage and enough authority to serve as a "single source of truth". This prevents different versions of different glossaries being around and causing confusion.
- Governance with respect to the thesaurus and its maintenance is well-organized. Ownership lays with the business, and processes are in place to keep the thesaurus up to date on a regular basis. This fosters innovation and successful IT projects.

In the next section we discuss standards for publishing a thesaurus using Web-technology.

[20] The information in this paragraph was kindly provided by Mr. N. van Gemmert, senior operative at the department of ASK and chief editor of the dictionary (personal communication).

Scoping and Sizing

An important practical question concerning the enterprise thesaurus is size: how many entries should be included so as to reach adequate coverage? Each organization must find its own answer. However, there are some general considerations. One has to find a balance between "low cost, high impact" and "high cost, low impact". To find this balance, it is useful to look at how lexicographers decide on the size of a dictionary.

It is well-known that the frequency distribution of words (or rather, lexemes) follows a so-called Zipf-distribution: the frequency of the second-most frequent word is about half the frequency of the most frequent word. The frequency of the third-most frequent word is about one third the frequency of the most frequent word, and so on (see Figure 4).[21]

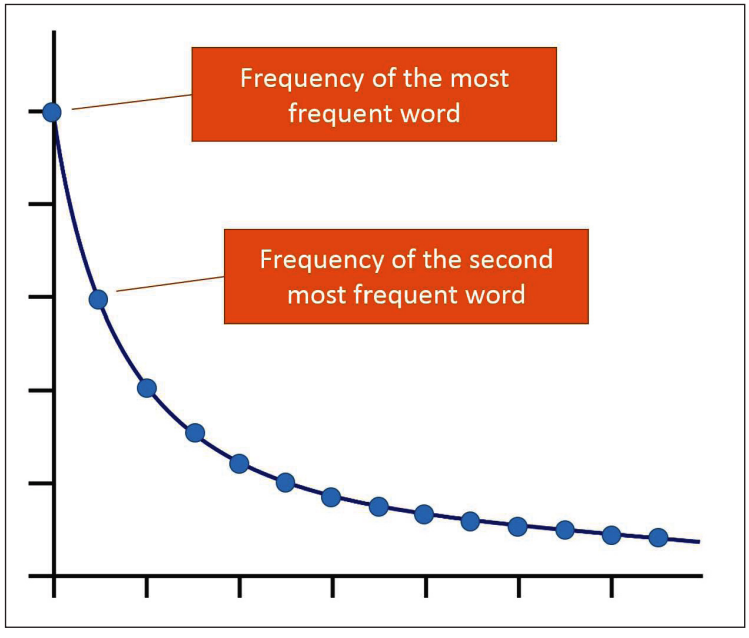


Figure 4. A Zipf-distribution of word frequencies.

This means that vocabulary of quite limited size suffices to understand a very large set of utterances. Recent statistical research in Dutch shows that knowing 5,000 words suffices to understand 95 percent of any Dutch corpus. To understand the remaining 5 percent of the text, you need to know some 200,000 additional words.[22]

This is reflected in the different types of dictionaries available on the market. Differentiating by number of entries, one finds on the high end of the spectrum the famous *Woordenboek der Nederlandse Taal* (Dictionary of the Dutch Language). The work was started in 1830 by Matthias de Vries. During one and a half century, five generations of editors worked on this dictionary which

[21] See Baayen (2002), Lauder (2010), and Pusted (2004).
[22] See Tiberius and Schoonheim (2014).

now contains more than 300,000 entries.[23] The dictionary is freely available online,[24] but its uptake is mainly restricted to scientific and linguistic circles. This means high cost, and, at least in terms of number of users, low impact.

On the other end of the spectrum we find pocket dictionaries with about 5,000 entries. These have a much broader audience. Moreover, for most general purposes a pocket dictionary suffices, given the result mentioned above that 5,000 words suffice to understand 95 percent of any corpus. Pocket dictionaries are often created by one single editor, sometimes a small team, and sold to a wide audience. This means low cost, high impact.

Therefore, when compiling an enterprise thesaurus, it is generally a good idea not to strive for completeness. Finding the right balance between cost and impact is not an exact science. One does not have to perform corpus statistics to find out what the significant terms are. Like in compiling an ordinary pocket dictionary, the expert opinion of the editor and his team should suffice to arrive at a sensible selection of entries, given purpose and budgetary constraints.

Governance

An effective enterprise thesaurus is authoritative. This makes it an instrument of language policy. It is a good practice that the functions responsible for the structure of the primary processes are also responsible for the content of the enterprise thesaurus. This is usually difficult to realize, because many concepts are fundamental to more than one process. The situation resembles a familiar problem well-known in relation to business process reengineering: the hierarchical lines of delegating responsibility define just one tree structure, but sometimes you need more. When two processes in different business units depend on each other, or both depend on the normative definition of the same term, one often has to rely more on cooperation than on flow of control or authority.[25]

One should therefore look for practical solutions. A mandating structure often works best. Preferably, in view of the multidisciplinary nature of the decisions to be made, the editors operate under the supervision of a board at the tactic or strategic level located close to policy making rather than that of (the manager of) a department or business unit. The supervisory board receives explicit mandate from senior executive management to ensure organization-wide adoption of the results.

Different temporary working groups with specialists can be assigned to address particular topics. It helps to have a small team of editors, maybe just one, who knows his way in the necessary tooling and has some experience in compiling glossaries and thesauri.

An important aspect of governance is continuity. Language changes constantly, and this is also – perhaps one should say especially – the case for organizational jargon. New laws and policies, new technologies and other

[23] See Schoonheim (2013)
[24] See http://wnt.inl.nl/.
[25] See for instance Smallwood (2014), Chapter 6.

innovations change the shape of the work and introduce new terms, and old terms become obsolete.

Maintenance of an enterprise thesaurus is a continuous process. Part of the governance structure is therefore a release schedule. For instance, one could decide on three minor releases per year and one major release. The minor releases are approved by the supervisory board, the major release by the senior executive board. The optimal frequency depends of course on multiple factors. The important point is that one needs to take continuous change into account.

Skills

In general, specific jargon terms are far easier to describe than high-frequency terms in general language. Some basic knowledge of elementary lexicographic practices will certainly come in handy. The most important topics include the different ways to describe word meaning, methods to deal with polysemy and insight in the nature of semantic relations.[26]

However, anyone with a deep understanding of the organization's processes and an above average feeling for language would qualify to contribute definitions to the enterprise thesaurus. It is not uncommon that people with a legal background end up making a large number of contributions. This is understandable because legal issues revolve around subtleties of language. There is also a risk involved in this: legally sound definitions are most often difficult or at least cumbersome to grasp for users without a legal background. Legal speak is to be avoided. The definitions in the thesaurus should be formulated in the language of the business, so that business persons are optimally supported in their work.[27]

5. Leveraging the web: linked data and SKOS

The benefits that an enterprise thesaurus has to offer can only be fully realized when the thesaurus is published on the Web – either behind a firewall for use inside the organization or, if desired, on the World Wide Web. In doing so, it is a good idea to observe the SKOS-standard. This section introduces the standard and discusses the practical benefits it offers.

SKOS as a standard

Simple Knowledge Organization System (SKOS) is a World Wide Web Consortium (W3C) recommendation designed for the representation of thesauri and similar controlled vocabularies.[28] Its roots go back to 1979, but the standard was formally released in 2008. Thesauri represented in SKOS

[26] A useful practical introduction is provided by Atkins and Rundell (2008).

[27] For a description of the skills required for compiling a thesaurus taking a library sciences perspective, see Hedden (2010), Chapter 2. Formal quality criteria for measuring thesauri have been described in Suominen and Mader (2014) and Mader and Haslhofer (2013).

[28] See Miles and Bechofer (2008). For discussion of the standard, its background and its application, see Baker et al. (2013).

are machine-readable and can be exchanged between software applications and published on the World Wide Web.

SKOS is based on the Semantic Web standards RDF (Resource Description Framework), RDFS (Resource Description Framework Schema Language) and OWL (Web Ontology Language), ratified by the W3C in 2003. These standards taken together are often referred to as Linked Data standards and have laid the groundwork for a major shift in the way we treat data and create IT systems.[29]

These standards are founded on a specific variant of referential semantics called *description logic*. As a consequence, an important point of departure in these standards is the requirement that symbols refer uniquely to "resources" (objects, things, concepts – essentially everything in the universe). To ensure that a symbol has a unique reference even when used across the Web, the standards prescribe the use of Uniform Resource Identifiers (URIs). An example is an email address, which, based on conventions and rules underlying the use of URI's, is guaranteed to uniquely refer to a unique email account.

A URL (Uniform Resource Locator) is a special case of a URI, namely, one referencing a resource on the Web, such as an HTML-document served by a Web server. Therefore, a URL is also called a Web address. A browser can use the URL of a resource on the Web to retrieve the resource and display it.

A URI, on the other hand, can also be defined – or *minted* – so as to refer to a resource *not* on the Web. Take for instance James Bond, the fictional British Secret Service agent. MI6 could decide to mint the URI http://data.mi6.gov.uk/JamesBond as a symbol that uniquely refers to the hero. Typing this URI in the address bar of your browser would then result in a 404-not found error message,[30] but this does not prevent one to use this symbol unambiguously and across contexts in different datasets on and off the Web to make statements about him. The same conventions and rules that guarantee that an email address refers uniquely to a single email account, also guarantee that the URI http://data.mi6.gov.uk/JamesBond refers uniquely to James Bond.

The fundamental element of the SKOS vocabulary is the concept. Concepts are the units of thought – ideas, meanings, or (categories of) objects and events – which underlie many knowledge organization systems. As such, concepts can be thought of as abstract entities which are independent of the terms used to label them. Once a concept is identified, various properties and relations can be assigned to it. These properties include prefLabel, indicating the preferred label for the concept – one for each language the thesaurus supports –, and altLabel, indicating alternative labels. The most significant relations are broader and narrower, following the ISO-standard discussed in Section 3 above.

[29] For a short non-technical introduction to Linked Data with a focus on concrete benefits, see Voskuil (2014).

[30] In fact, it is considered good practice that the URI's host (in our example, data.mi6.gov.uk) returns "useful information" about the resource not on the Web when a browser tries to resolve such a URI. As long as one keeps in mind that this is just information about the resource being referenced, the convention should not lead to confusion. See Wood (2010) for detailed discussion and further references.

In SKOS, concepts are identified by URI's. As a result, an organization must mint URI's to uniquely refer to the concepts it defines. This has the advantage that concept identifiers are globally unique. Suppose that the concept of capacity as defined by DJI, discussed in Section 4 above, would be given the identifier http://vocabulary.dji.nl/Capaciteit. One could then use this identifier in any context, and it would be clear immediately that it is the DJI-concept that is meant – and not some other concept of capacity.

Also relations in SKOS are identified by URI's. Take for example the SKOS relation referenced by http://www.w3.org/2004/02/skos/core#broader, which, using standard conventions, can abbreviated as skos:broader. Whenever one encounters this URI, it is clear that what is meant is the relation "broader" as defined in SKOS. The interpretation of the symbol remains constant across contexts.

This constancy of meaning across contexts is useful on the Web, because information expressed by such symbols can be exchanged between no matter how many IT systems while the semantics of these symbols remains constant and explicit, without the need of making additional arrangements.

Importantly, SKOS expresses the same conceptual structures we already encountered when introducing thesauri in Section 3 above. In addition, SKOS ensures that these structures are machine-readable in a standardized way. Let us briefly consider an example.[31] In the following, each expression containing a colon represents a URI. For the linguistically inclined: the first line is a sentence of the form subject-predicate-object, the indented lines are coordinated sentences of the same form with ellipsis of the subject:

```
ex:animals rdf:type skos:Concept;
    skos:prefLabel "animals"@en;
    skos:altLabel "creatures"@en;
    skos:prefLabel "animaux"@fr;
    skos:narrower ex:mammals.
ex:mammals rdf:type skos:Concept;
    skos:definition "Mammals are endothermic amniotes distinguished from
    reptiles and birds by the possession of hair and..."@en;
    skos:prefLabel "mammals"@en;
    skos:broader ex:animals.
```

The above SKOS-statements declare that ex:animals is a resource of the type Concept as defined in SKOS. It has the preferred label "animals" in English and "animaux" in French, plus an alternative label "creatures" in English. Furthermore, it has a narrower term, namely, the concept ex:mammals. The latter concept has a preferred label and a definition in English and also has a broader term, namely, the concept ex:animals. The above statements are machine readable, but not friendly for human readers. To see why it is so

[31] The example is taken from Isaac and Summer (2009) in a slightly modified form.

useful to capture a thesaurus using SKOS statements, we need to dig deeper in the way these statements can be used on the Web when published.

Publishing and connecting thesauri

To create, maintain and publish and present SKOS-based thesauri one uses tools, of which there are many available.[32] Such tools offer specific support for the authoring and maintaining a thesaurus by offering user-friendly functions to editors. In addition, such tools offer support for publishing the thesaurus on the Web. This is done in two forms: machine readable in the form of SKOS statements (see above), and human readable in the form of a navigable Web application. Figure 5 and Figure 6 show screenshots from one of the available tools.[33] Figure 6 shows a page from the freely accessible Wolters Kluwer Deutschland Thesaurus of Labor Law.[34]

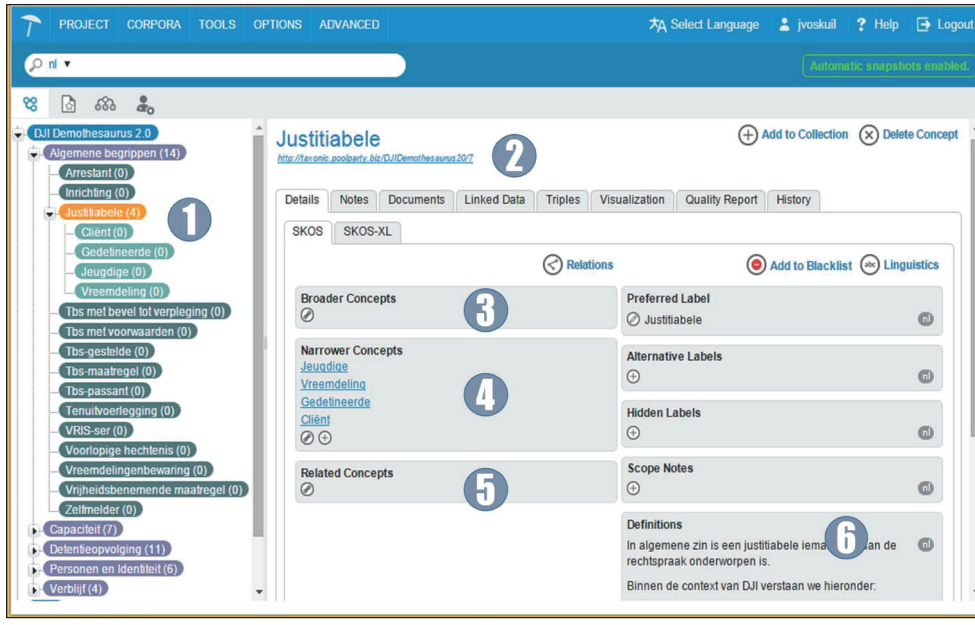


Figure 5. A screen for editing a thesaurus. The left pane (1) shows a hierarchical tree structure representing the thesaurus structure. The right pane shows the details for the selected concept (2). There are boxes for adding broader (3), narrower (4) and related terms (5), definitions in different languages (6), and so on.

[32] The Wikipedia article on SKOS presents an overview of such tools. See http://en.wikipedia.org/wiki/Simple_Knowledge_Organization_System.

[33] The screenshots in Figure 5 and Figure 6 show screens generated by PoolParty, a widely used thesaurus management tool created by the Vienna-based software vendor SWC (Semantic Web Company). Permission to use these screenshots was kindly provided by Mr. A. Blumauer of SWC and Mr. C. Dirschl of Wolters Kluwer Germany.

[34] See http://vocabulary.wolterskluwer.de/arbeitsrecht.html.

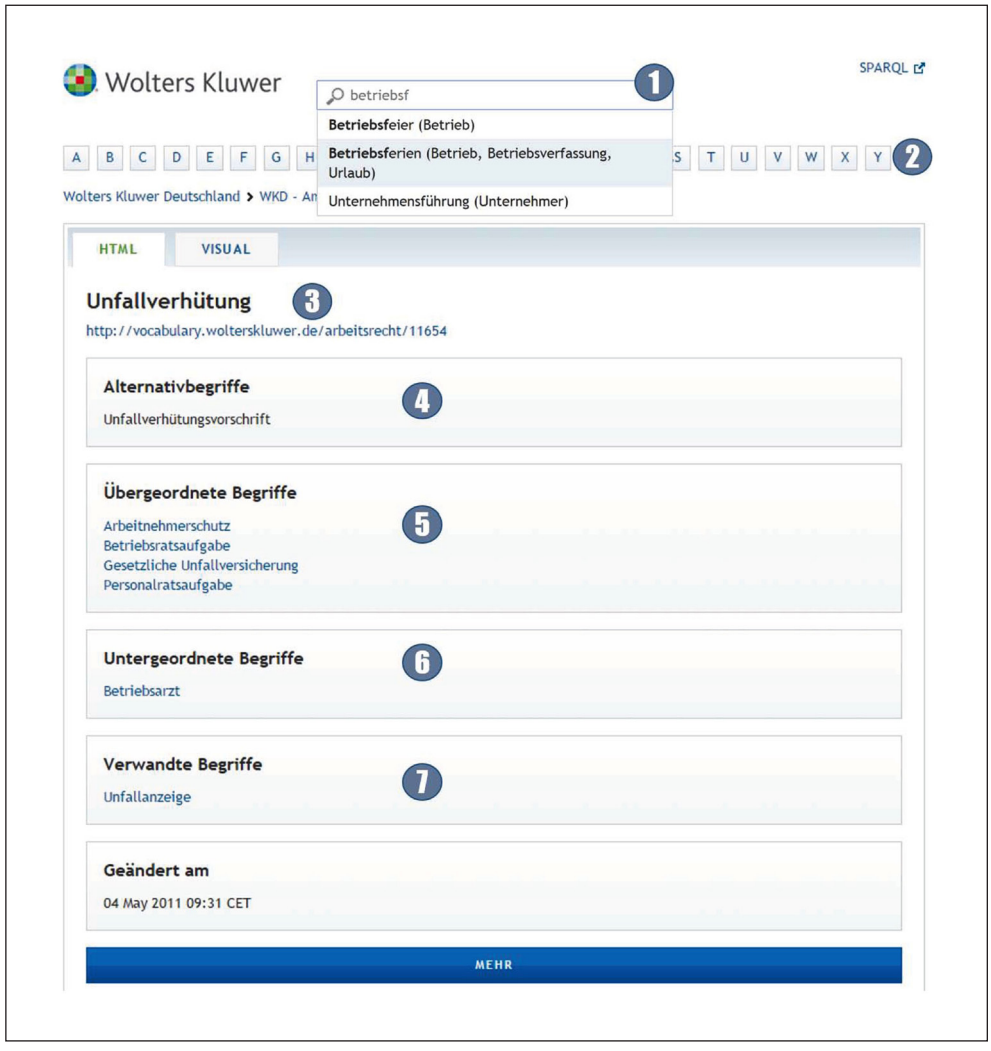


Figure 6. The end-user view. The page provides a search function with auto-complete (1) and a function for browsing terms alphabetically (2). The page shows information about the concept with the prefLabel Unfallverhütung (Accident prevention) (3). Directly below the prefLabel, the concept's URI is printed. The rest of the page shows (among other things) alternative labels, broader terms, narrower terms and related terms ((4) through (7)).

In addition to functions for authoring and publishing thesauri, most tools offer functions for using thesauri in information retrieval, such as services for tagging content, faceted search and so on.

The availability of SKOS-compliant thesauri on the Web - the number of these is growing rapidly - creates ample opportunities for connecting content from other sources. An example is given in Figure 7. It shows the bottom part of the same page from the Wolters Kluwer Labor Law Thesaurus we saw in Figure 6 above. The upper part of the page, in Figure 6, shows information from the Wolters Kluwer thesaurus itself. The bottom part, shown in Figure

7, shows information from other sources such as DBpedia and the Eurovoc thesaurus. For a thorough discussion of the technology used in creating the thesaurus as well as a description of its objectives and the consequences of semantic linking for business models in the publishing world, see Dirschl et al. (2014).

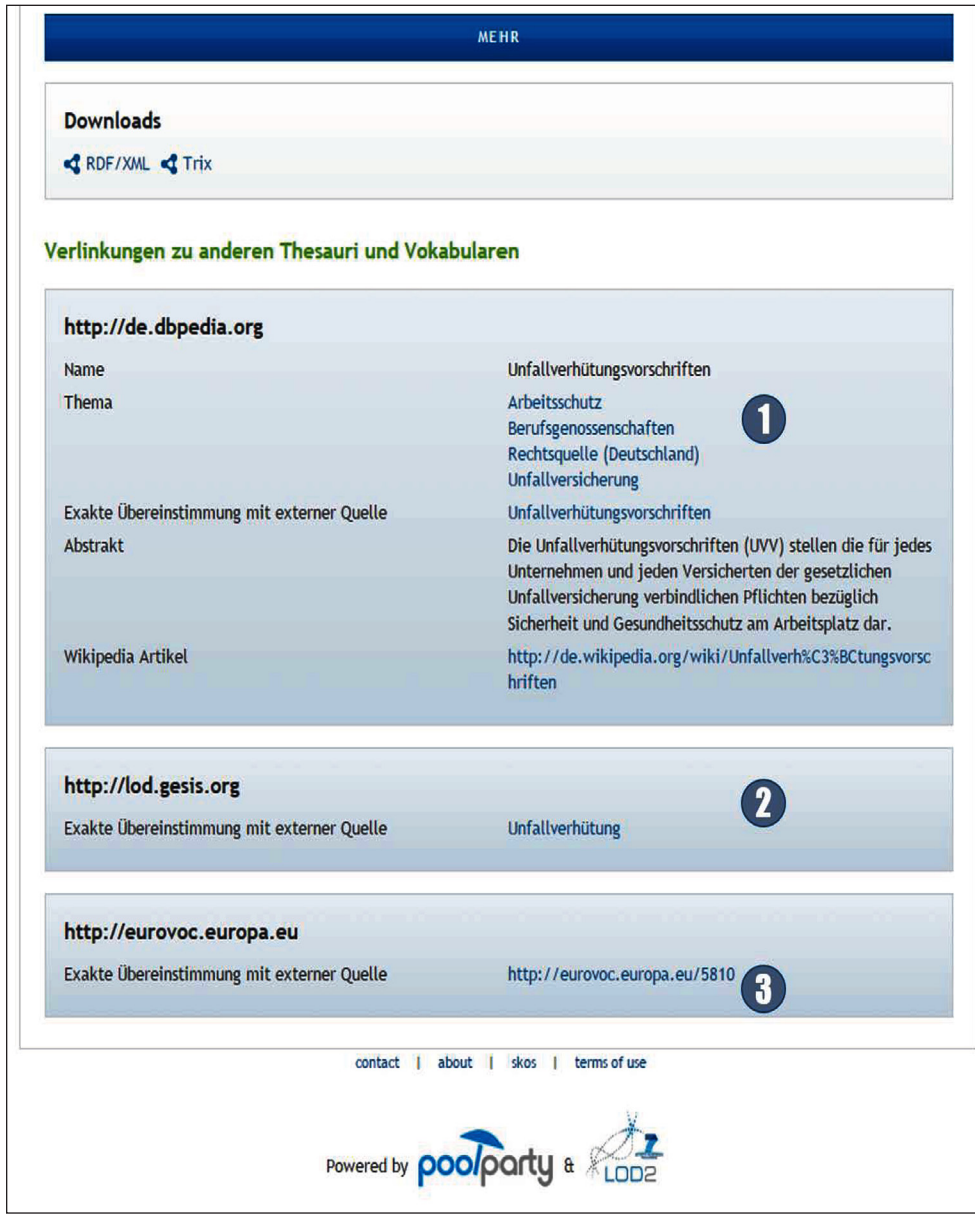


Figure 7. The bottom-part of the page showing information about "Unfallverhütung". It shows a definition from DBpedia with a link to information in Wikipedia (1) and provides links to the exact matching terms in the Gesis-thesaurus (2) and the Eurovoc-thesaurus (3).

Two important points deserve our attention. First, the information from others sources shown on the page alongside the publisher's own information is retrieved real-time, at the moment the user requests the page. The traditional way of integrating information from different sources would be to export data from an outside source and importing the data in the publisher's systems, which includes transforming them to match a particular data format.[35] Web-technology in combination with semantic standards enables information integration at a fraction of the cost.

Second, the linkage between the thesaurus and the external information is done on a conceptual level, and is not just a matter of a hyperlink to a piece of text. This can be illustrated by comparing Wikipedia to DBpedia. Wikipedia is essentially a collection of articles, each of which treats a separate topic. It does not cater for a function to generate lists such as a list of all popes in the period 1900 – 1970, or a list of the rivers that tribute to the Rhine and are more than 100 kilometers long. DBpedia, on the other hand, is specifically designed to be able to generate such lists, based on Linked Data standards.[36]

Based on the same technology, SKOS makes it possible for example to incorporate a list of all narrower terms of a given concept from one thesaurus into another. This offers a much deeper level of connecting and integrating information.

The ability to enrich a thesaurus by adding information from different sources on the Web offers several concrete use cases in the context of creating enterprise thesauri:

- The definition of a term can be directly compared to definitions used in other organizations with which information is exchanged.
- Hyperlinks can be generated for giving direct access to sources of background information, such as Wikipedia or legal resources.[37]

Publishing the enterprise thesaurus on the Web in conformance to SKOS makes the thesaurus machine-readable. This also offers several concrete use cases. To mention some examples:

- The content of the thesaurus can be used in other systems, such as in help texts in processes support systems.
- Semantic relations in the thesaurus can be directly used in application logic. For instance, when the user of a system must select a value for the

[35] Exporting and importing is how the Justitiethesaurus, developed and maintained by WODC (Wetenschappelijk Onderzoek- en Documentatiecentrum) - the scientific institute of the Dutch Ministry of Security and Justice - is reused by other organizations, such as the Council for Child Protection. Personal communication by C.J. van Netburg, editor of Justitiethesaurus. Work on this thesaurus started before the advent of SKOS, which explains why it is distributed in a proprietary format. See Van Netburg (2013).

[36] See Mendes, Jakob, and Bizer (2011).

[37] Interestingly, initiatives are underway in eGovernment to make legislation directly available on the Web at a granular level. An example is LiDO, a program run by the Dutch government, see https://data.overheid.nl/english. A general approach using LD for legal content is described in Hondros (2010) and Santosuosso and Malerba (2014). For a thorough analysis of the state of the art in Europe and future developments, see Van Opijnen (2012, 2014).

field "country of origin", the drop-down menu listing the options could be directly taken from the thesaurus, which then acts as a repository of master data.[38]

- Documents such as work instructions or template letters, when presented to a user on a computer screen or a tablet, can be enriched automatically with hyperlinks to the thesaurus, so that term definitions can be shown in tooltips.

In reality, the possibilities go far beyond the concrete uses cases described here. These possibilities are not completely new: by using pre-Web information technology, the same results can be obtained, but only at prohibitive costs. By taking away this economic barrier, SKOS offers a completely new playing field for organizations to manage their vocabularies and take knowledge management and language policy to the next level.

Linking thesauri to IT artifacts

We now return to the theme introduced at the beginning of this paper: the use of the enterprise thesaurus in the interaction between business and IT suppliers. The previous subsection describes how information in thesaurus can be used directly in IT systems. The thesaurus also plays an important role in process of system development itself: it provides the basic concepts, their definitions and semantic structure for the system design process to build upon.

Taking this one step further, the relation between a class or object in a design artefact and the corresponding concept in the thesaurus can be made explicit. In its simplest form, this can be realized in a textual form. Better still is to use direct hyperlinks in the design documentation. Conversely, links to the design documentation can be added to the relevant concepts in the thesaurus, so that navigation is supported both ways.

Ideally, when the systems under development are themselves based on Linked Data standards, the concepts in the design and in the thesaurus are directly interlinked. For instance, in many of the normative standard vocabularies ratified by the W3C, design concepts are treated as SKOS concepts. Consider the following schema (Figure 8) taken from Registered Organization Vocabulary, introduced above.[39]

[38] Increasingly, standard bodies publish controlled vocabularies using SKOS. See the next subsection for an example.

[39] See Archer, Meimaris, and Papantoniou (2013).

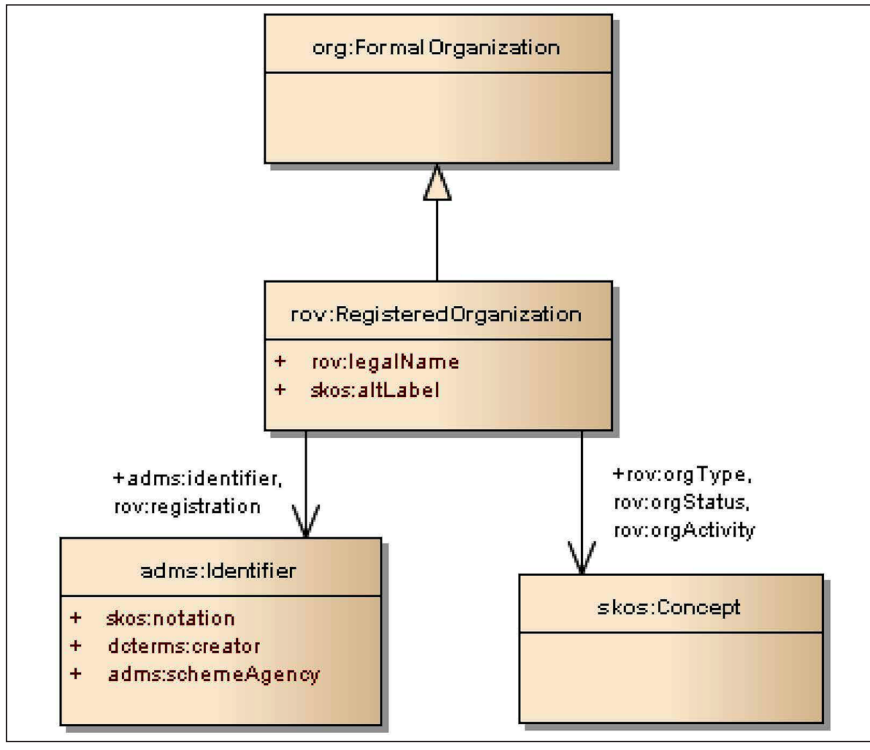


Figure 8. Schema representing major resources and properties in the Registered Organization Vocabulary.

The schema says, among other things, that the concept of registered organization has a legal name and a skos:altLabel, in other words, an alternative label as defined in SKOS. Thereby the concept of registered organization itself is treated as a skos:Concept. It can be incorporated directly in the enterprise thesaurus, or linked to a concept in the enterprise thesaurus using a relation like skos:exactMatch or owl:sameAs.

Note, incidentally, that the schema presents three semantic relations linking a registered organization to a skos:Concept. This shows that the possibility of using a thesaurus a source for master data, mentioned in the previous subsection, is actually practiced. For instance, the relation rov:orgActivity indicates the economic activity the organization performs. This activity is expressed as a skos:Concept, hence, a concept in a SKOS compliant thesaurus. The standard states that the preferred choice for European interoperability is NACE (Nomenclature statistique des activités économiques dans la Communauté européenne), the SKOS-based thesaurus of economic activities published by Eurostat.[40] For a more elaborate discussion of the role of SKOS and Linked Data in the enterprise IT-landscape, see Izza, Vincent, and Burlat (2006).

[40] See http://ec.europa.eu/eurostat/ramon/nomenclatures/index.cfm?TargetUrl=LST_NOM_DTL& StrNom=NACE_REV2&StrLanguageCode=EN.

6. Reflections

The enterprise thesaurus as a strategic information hub

Summarizing, in the transition from a problem statement in the business to delivery of an IT-system, the enterprise thesaurus plays a crucial role. It provides definitions and description of usage of terms. A restricted amount of simple, intuitive semantic relations between terms add value. The abstract nature of these relations caters for flexible usage, for instance to indicate class membership, the theme-subtheme relation or the relation between a feature and feature values.

When all the potential benefits of the enterprise thesaurus are realized, the thesaurus starts to function as the organization's central knowledge hub. To business persons, IT analysts, and stakeholders outside the organization, the thesaurus acts as a knowledge base where the language of the business is the structuring principle. It is the concepts, the terms used to refer to them and relations between concepts that provide the entry points into and navigation paths through this body of knowledge.

Within this structure, definitions of terms and their explanations are combined with and linked to other concepts, other thesauri, legislation and compliancy literature, IT design documentation, sources of background information from inside and outside the organization, work instructions, textual templates, and so on: there are no real limitations - at least no technical ones. Thus, the enterprise thesaurus is an important tool for knowledge management.

Still, it is important that governance is well established. The enterprise thesaurus must be normative to be able to fulfil its basic function of being a reliable source of truth. This means, first and foremost, that the thesaurus should be of a reasonable size. As we have seen, what counts as reasonable will be different for different organizations and depends on purpose and budget. As a rule, one should strive to find a balance between just large enough to have reasonable coverage, and small enough to keep matters manageable from a governance point of view. The principle of preference is: low cost, high impact.

Once a normative, broadly used and manageable SKOS thesaurus is in place, adding more and more links to it can turn the thesaurus into the centrepiece of a strategy to gradually adopt Linked Data in the long-term. Seen from this perspective, SKOS offers a low barrier method to introduce Linked Data in the enterprise – one which immediately yields concrete benefits without impacting existing IT assets.

A cautionary note is order, though. A knowledge base defines terms but still leaves open discussion about their application. Suppose someone dies because of an overdose of drugs. Was it murder, suicide or an accident? This is a question of judgment. However, clear definitions will help making the decision transparent.

What does this mean for lexicography?

In an article on the future of dictionaries, the director of Dictionaries of Oxford University Press Judy Pearsall writes that the dictionary of the future contains a new type of lexical content, namely, one that is structured and semantically annotated in such a way that it can be read intelligently by a machine, and new products, links and information are automatically produced as a result. It starts with a simple hub, but supports links to additional content to be added incrementally.[41] New business models need to be developed to keep lexicographic content creation an economically healthy enterprise. Work by Dirschl et al. (2014) in connection to the open legal thesaurus published by Wolters Kluwer introduced in Section 5 indicates the direction in which this may go.

The technological trends described in this paper with respect to the enterprise thesaurus can easily be projected on general lexicographic practice. However, lexicography in an organizational context focuses on jargon and highly domain specific terms and meanings. Applying the techniques to lexicography in the general domain requires the development of new Linked Data standards that are specific to lexicography, in the same way as SKOS is specific to knowledge management. A step in that direction is the formation of the W3C Ontology-Lexicon Community Group in 2012.[42]

References

Archer, P., M. Meimaris, and A. Papantoniou. 2013. "Registered Organization Vocabulary". [W3C Working Group Note 01 August 2013; http://www.w3.org/ TR/vocab-regorg/.]

Atkins, Beryl T. Sue and Michael Rundell. 2008. *The Oxford guide to practical lexicography*. Oxford: Oxford University Press.

Baayen, R. H. 2002. "Word Frequency Distributions", *Text, Speech and Language Technology* 18: 13ff.

Baker, T., S. Bechhofer, A. Isaac, A. Miles, G. Schreiber, and E. Summers. 2013. "Key choices in the design of Simple Knowledge Organization System (SKOS)", *Web Semantics*; *Science, Services and Agents on the World Wide Web* 20: 35-49. [Http://www.sciencedirect.com/science/article/pii/S1570826813000176.]

Bursch, Johannes. 2011. "Corporate Language Management at Daimler AG; Role and challenges". [Paper, META Forum, Budapest, 27-28 June; Available at http://www.meta-net.eu/events/meta-forum-2011/talks/johannesbursch.pdf.]

Dardjowidjojo, Soenjono. 1998. "Strategies for a successful national language policy: the Indonesian Case", *International Journal of the Sociology of Language* 130: 35-47.

[41] See Pearsall (2013).

[42] See http://www.w3.org/community/ontolex/.

Dirschl, Christian, Tassilo Pellegrini, Helmut Nagy, Katja Eck, Bert Van Nuffelen, and Ivan Ermilov. 2014. "LOD2 for Media and Publishing", in: Sören Auer, Volha Bryl, and Sebastian Tramp (eds), *Linked Open Data - Creating Knowledge Out of Interlinked Data; Results of the LOD2 Project*, pp. 133-156. Berlin: Springer.

DJI. 2014. "Gevangeniswezen in Getal 2008–2013". [Https://www.dji.nl/Organisatie /Feiten-en-cijfers/.]

Fredriksson, Riikka, Wilhelm Barner-Rasmussen, and Rebecca Piekkari. 2006. "The multinational corporation as a multilingual organization; The notion of a common corporate language", *Corporate Communications; An International Journal* Vol. 11/4: 406-423.

Frischmuth, Philipp, Jakub Klímek, Sören Auer, Sebastian Tramp, Jörg Unbehauen, Kai Holzweissig, and Carl-Martin Marquardt. 2012. "Linked Data in Enterprise Information Integration", *Semantic Web* 0: 1-17.

Furnas, G. W., T. K. Landauer, L. M. Gomez, and S. T. Dumais. 1987. "The vocabulary problem in human-system communication; An analysis and a solution", *Communications of the ACM* Vol. 30/11 (November): 964-971.

Hedden, Heather. 2011. *The accidental taxonomist.* Medford, NJ: Information Today.

Hondros, Constantine. 2010. "Standardizing legal content with OWL and RDF", in: David Wood (ed.), *Linking Enterprise Data,* pp. 221–241. New York: Springer.

Isaac, A. and E. Summer. 2009. "SKOS Simple Knowledge Organization System Primer". [W3C Working Group Note 18 August 2009; http://www.w3.org/TR/2009/NOTE-skos-primer-20090818/.]

Izza, S., L. Vincent, and P. Burlat. 2006. "A framework for Semantic Enterprise Integration", in: D. Konstantas, J. P. Bourrières, M. Léonard, and N. Boudjlida (eds), *Interoperability of enterprise software and applications*, pp. 75–86. London: Springer London.

Lambe, Patrick. 2007. *Organising Knowledge; Taxonomies, Knowledge and Organisational Effectiveness*. Oxford: Chandos.

Lauder, Alan F. 2010. "Data for lexicography; The central role of the corpus", *Wacana, Journal of the Humanities of Indonesia* (Lexicon and semantics) 12/2: 219-242.

Mader, Christian and Bernhard Haslhofer. 2013. "Perception and relevance of quality issues in web vocabularies", in: Marta Sabou, Eva Blomqvist, Tommaso Di Noia, Harald Sack, and Tassilo Pellegrini (eds), *Proceedings of the 9th International Conference on Semantic Systems (I-SEMANTICS '13)*: 9-16.

Mendes, P. N., M. Jakob, A. García-Silva, and C. Bizer. 2011. "DBpedia spotlight: shedding light on the web of documents", in: Chiara Ghidini, Axel Cyrille Ngonga Ngomo, Stefanie Lindstaedt, and Tassilo Pellegrini (eds), *Proceedings of the 7th International Conference on Semantic Systems, I-Semantics 11*, pp. 1–8. New York: ACM.

Miles, A. and S. Bechhofer. 2008. "SKOS Simple Knowledge Organization System Reference". [W3C Recommendation; http://www.w3.org/TR/

skos-reference/.]
Netburg, C.J. van. 2013. "Justitiethesaurus 2013; Gestructureerde standaard trefwoordenclassificatie". [WODC; https://www.wodc.nl/images/justitiethesaurus-2013_tcm44-531500.pdf .]
Opijnen, Marc van. 2012. "The European Legal Semantic Web; Completed Building Blocks and Future Work", *European Legal Access Conference*. [Http://ssrn.com/abstract=2181901.]
Opijnen, Marc van. 2014. *Op en in het Web; Hoe toegankelijkheid van rechterlijke uitspraken kan worden verbeterd.* PhD thesis, University of Amsterdam.
Paauw, S. 2009. "One land, one nation, one language; An analysis of Indonesia's national language policy", in: H. Lehnert-LeHouillier and A.B. Fine (eds), *University of Rochester Working Papers in the Language Sciences* 5(1): pp. 2-16. Rochester, NY: University of Rochester.
Pavort, Anna. 2005. *The naming of names; In search for order in the world of plants.* London: Bloomsbury.
Pearsall, Judy. 2013. "The future of dictionaries", *Kernerman Dictionary News* 21 (July): 2-5. [Http://kdictionaries.com/kdn/kdn21-1.html.]
Pusted, R. 2004. "Zipf and his heirs", *Language Sciences* 26/1: 1-25.
Santosuosso, A. and A. Malerba. 2014. "Legal interoperability as a comprehensive concept in Transnational Law", *Law, Innovation and Technology* 6/1: 51-73. [Http://www.ingentaconnect.com/content/hart/lit/2014/00000006/00000001/art00003, Http://www.unipv-lawtech.eu/files/SantosuossoMalerba-LIT-Interoperability.pdf .]
Schoonheim, T. 2013. "A brief account of Dutch lexicography", *Kernerman Dictionary News* 21 (July): 16-23. [Http://kdictionaries.com/kdn.]
Smallwood, R.F. 2014. *Information governance; Concepts, strategies, and best practices*. Wiley.
Steinhauer, Hein. 2005. "Colonial history and language policy in Insular Southeast Asia and Madagascar", in: Alexander Adelaar and Nikolaus P. Himmelman (eds), *The Austronesian languages of Asia and Madagascar,* pp. 65-85. London: Routledge.
Suominen, O. and C. Mader (2014). "Assessing and improving the quality of SKOS vocabularies", *Journal on Data Semantics* 3(1): 47-73. [Https://eprints.cs.univie.ac.at/3707/.]
Tiberius, C. and T. Schoonheim. 2014. *A frequency dictionary of Dutch; Core vocabulary for learners.* Taylor and Francis.
Voskuil, J. 2014. "Linked Data in the Enterprise – Second Edition". [A Taxonic Whitepaper; http://taxonic.com/white-papers.]
Voskuil, J. 2014. "Vocabulary management and SKOS; Putting the business in the lead". [Paper, SEMANTiCS 2014, Leipzig, 5 September.]
Wood, David. 2010. "Reliable and Persistent Identification of Linked Data Elements", in: David Wood (ed.), *Linking Enterprise Data*, pp. 149-172. New York: Springer.