Evolving and Exploiting Domain Terminologies

Thomas Bandholtz¹, Joachim Fock²

Abstract

Domain terminology is a network of hopefully shared conceptualisations in the area of discourse, designated by hopefully un-ambiguous terms and labels or - in terms of data representation - identifiers. In most cases those terms and identifiers are not managed on the enterprise level which causes misunderstandings and mismatches in all areas of communication. This article drafts a roadmap of evolving and exploiting domain terminologies based on experiences of the authors and further thoughts about the next steps to go.

1. Introduction

A growing number of organisations have appreciated the role of a formalised domain terminology which has been negotiated and agreed to by all relevant stakeholders. They are utilising such terminologies in their information infrastructure or preparing to do so. This article is based on the author’s experience from appropriate projects since the 1990s, but in parts we will go beyond what has been implemented to draft a comprehensive scenario of the near future.

We start with constructing an initial terminology from scratch or based on existing “semantic assets” and exemplify several useful applications from knowledge management to linked data integration.

In this scenario, human intellectual work goes into a close coalition with agents implemented by machines. Not every described feature has been integrated in the course of these projects, but most of them have been implemented somewhere and may be integrated and extended.

Exploring terminology as “links between objects, concepts, and their terminological representations” we refer to (ISO 704:2009) “Terminology work” and, more specific, to the W3C Recommendation “Simple Knowledge Organisation System” (SKOS, 2009).

2. Setting up and Maintaining Terminologies

A domain terminology is nothing that you will have to invent. It is already there, spread all over the document corpus and the charivari of every day’s face-to-face discussions and conference calls. A well-built terminology is “true” in a sense of a true representation of the common wording patterns, but it will try to make those patterns more explicit, targeting to minimise existing babel and noise. We will see how it works in section 3 and take a look at the terminology construction at first. It is simply a fact that you will have a terminology at hand before you can start to apply it, but the intended use cases are setting urgent priorities (“competency questions”) for the construction, of course. So you may even read section 3 first if you prefer to have a clearer focus on requirements before you start to build.

¹ innoQ Deutschland GmbH, Krischerstr. 100, 40789 Monheim am Rhein, Germany. thomas.bandholtz@innoq.com
² Umweltbundesamt, FG I 1.5, Wörlitzer Platz 1, 06813 Dessau-Roßlau, Germany. joachim.fock@uba.de
Here we describe a semi-automated procedure which starts right from scratch. If your organisation already maintains some established glossary or thesaurus the same procedure can provide quality assessment and systematic extrapolation – language never sleeps, while thesauri obviously may do so: GEMET, f.i., has last been edited in 2000 and does know neither “emission trading” nor “nuclear power phase-out”.

2.1 Existing Approaches

Someone will have done some work for you before, maybe not knowing. There are product catalogues, sitemaps, thousands of code lists, and maybe even some outdated glossary. Maybe you have some domain standards such as the Financial Industry Business Ontology (FIBO) in the finance world.

Take all you can find, and then select approximately 100 most significant terms. This restriction is a good trick that helps provoking a controversial discourse about terms and the concepts behind them – something the team will need to elaborate on.

2.2 Examining the Corpus

Here we need some 100 (again) documents that describe your organisation at large. Each document is linked to the organisational unit where it comes from. Now a machine agent generates a ranking of all terms in this corpus and compares this with the 100 terms selection from step 1.

Both lists now should be revised and merged to set up the skeleton of the terminology.
2.3 Concepts and their Denotations

In this state of work it should have become obvious that one concept may have multiple “terminological representations”[ISO 704:2009], such as “waste” and “garbage”. In SKOS a concept is “an idea or notion; a unit of thought”, while a label is is a string of UNICODE characters, such as ‘romantic love’ or "れんあい", in a given natural language, such as English or Japanese (written here in hiragana).”

It is up to you if “garbage” and “waste” should be two labels of a single concept or two different concepts as well. If your domain is waste management, you will probably want to distinguish, in other cases may be not. You will have to consider that distinguishing on a very fine granularity level will take many resources.

If you have multiple labels of concepts, one of those labels should be the preferred label, and the others alternate. SKOS also supports a hidden label, and you may add your specific types such as scientific or vernacular names.

2.4 Hierarchy (Granularity)

Some concepts are obviously wider and others are narrower in their meaning. When you start building hierarchies you will soon come to the question which concepts should make your top concept set up, the entry points (root nodes) to your concept tree. Together they should represent the whole area of discourse, and the semantic distance should be evenly distributed. Aligning this list is an intellectual challenge. Any algorithm to compute semantic distances is based on path length, and so far we only have a hierarchy. We can try to make each sub-tree behind these top concepts equal in size, but this is not mandatory. Professional thesauri use to distinguish three kinds of hierarchy:

1. Generic: wider and narrower meaning.
2. Partitive: the narrower concepts are parts of the wider: wheels and motor are parts of the car.
3. Instantive: a class-individual type of hierarchy: Micky is a cat.

If you consider using this, you have to make sure that you will have tool support.

2.5 Relations

Classical thesauri support generic non-hierarchical relations between concepts, and it may be a good idea to add more specific types of relations according to your domain. Examples may be “smoking causes cancer” or “car eats mouse”. If you want to do so, again you have to make sure that you will have tool support.

Working with typed relations will force you to decide whether smoking really cause cancer or is just “somehow” related. Someone will propose another relation type such as raises-the-probability-of, and the next speaker wants to add a percentage. If you really want to set up a terminology, you should avoid all this, at least at this stage. You won’t need it to support the most nearby use cases, and the risk of drifting into an endless debate about “ontology” has to be taken serious. There is nothing wrong with ontology, but we are focussed on something much simpler in this article. Once you have a stable (and maintainable) terminology, you may add any kind of statements about your concepts, but you will have a stable core concept tree.
2.6 Population

So far we have some hundred terms with some of them denoting the same concept, ordered by a concept hierarchy. But we also have a huge list of further terms that have appeared in the document corpus, and all of them are candidates to be integrated.

Machine agents can support this with statistics: frequency of each term, frequency of concurrent appearances of two terms in the same sentence/section. From here there may come some hints to rank relevance and similarity to established concepts.

The agent will not be able to decide whether two very close concepts have a hierarchical or some other kind of relation – this is something left for the team.

2.7 Definitions

In some cases concepts may be well-defined, but mostly there are multiple definitions coming from different sources with different viewpoints. You may link all these definitions to a single concept, or you may set up a dedicated concept for each viewpoint and link these concepts to a broader generic. This again is a question of the intended granularity (do you really need to distinguish? What will you miss when you don’t?).

Whatever you prefer, there is absolutely no problem with multiple definitions and not even with no definition at all. A concept may be solely represented by its label(s) and relations, so you do not need to force complete definitions before you deploy the terminology.

If you quote definitions, do not forget to refer to the source.

2.8 Lexical Representations

In the document corpus we just have character string hopefully matching some lexical representation of a concept. It is not only that concepts may have multiple labels, it is even worse: each label will have multiple lexical representations. The complexity in this area depends much on the language(s) to be supported. The “awful German language” has already been acknowledged by (Twain, 1880)3, but English has some lexical stumbling blocks as well. For example: “waste water” may also be written “wastewater”, and “wasted” may be a flexional form of the verb waste, but “wasted water” is slightly different from wastewater.

While the flexional forms are quite simple to handle in English, term composition remains a challenge.

2.9 Formalisation and Deployment

In order to be utilised in the following production, the terminology needs to be available in a formal, machine readable language. For widest interoperability we recommend SKOS, the S3C Recommendation. Now we need a Web application to provide both machine readable (RDF) and human readable (HTML) representations from the same source. Each concept will be represented by a unique HTTP URI that people can bookmark and use as a concept reference.

No matter whether you go public or restrict access to an intranet, make your terminology work in a Web architecture. If you prefer open source, you may consider iQvoc4 as an adequate tool.

3 http://www.crossmyt.com/hc/linghebr/awfgrmlg.html
4 https://github.com/innoq/iqvoc/wiki/about-iQvoc
2.10 Persistence and Evolution

“What makes a cool URI?
A cool URI is one which does not change.
What sorts of URI change?
URIs don’t change: people change them.”
(Berners-Lee, 1998)

Persistent URIs are required because we want people to use URI references when they refer to a concept, and those links should never break. If you consider deleting a concept, mark it “expired” instead because someone may link to it.

If you have a new version of concept add it like a new concept und link it to the expired predecessor. If you just have minor changes like correcting a misspelling in some definition, you don’t need a new version at all. We would not version the terminology as a whole, but we recommend a permanent, downwards compatible evolution. This may be in conflict with some intended use cases, but it is the best strategy if you want to keep your terminology alive.

3. Applied Terminology

There are many ways of exploiting a brave terminology, and here are some of them.

3.1 General Communication

The most obvious (but hardly measurable) benefit you will have from a well aligned and well deployed terminology will be a better understanding in everyday communication. People can refer to clarified concepts and misunderstandings can be avoided. To make this work, you will have to do some promotional tricks. You may elect a concept (or -non-concept) of the month and initiate a free discussion in some wiki, or you may comment on the wording of the most popular documents, and so on.

3.2 Knowledge Management

Conventional search approaches are based on a full text index and character string composition. Even if you handle the lexical representation well, you are stuck to comparing labels instead of concepts. For example, if you are searching for “garbage” you will never find documents that consequently use “waste”.

Some simple improvement is providing synonym lists in the search interface, where users may add terms just by picking them from this list.

More consequently you should maintain a concept-based index of the document corpus along with the full text index so people may search for exactly “garbage” (in the full text index) or for any denotation of the respective concept (in the concept-based index) without having to pick synonyms from a list.

3.3 Automated Indexing

If you want a concept based index you shouldn’t start to tag each document using concepts manually, at least not at large. Available tools (such as iQvoc, again) provide services to do this automatically with a reasonable quality – depending on the semantic and lexical “wealthiness” of your terminology. Especially detecting composed terms need some patient registration of such compositions. You will gain more benefit from working on your terminology than from manual indexing.
3.4 Retrieval

If you have automatic indexing, you can as well apply this procedure to the user’s search terms. This is especially useful when you encourage people to use complete document sections or even complete documents as search conditions – you can validate the topical similarity of documents (not the similarity of the content positions).

This is a new retrieval pattern that users will have to get used to. But if you find the most similar documents whenever a new document arises, you can provide links to these similar documents anytime a user comes across any single document.

3.5 Push Services

As we have persistent URIs for each concept, concepts can be used just like points-of-interest in spatial navigation. Users may subscribe to concepts-based notification services with several options. Those services may scan news feeds for “all oil company names or direct relations of the concept oil” for example and sent you the matching feed selection if you care for some oil company share.

3.6 Linked Data

The most sophisticated exploitation of terminology comes with linked data. The four basic principles are:

1. Use URIs as names for things
2. Use HTTP URIs so that people can look up those names.
3. When someone looks up a URI, provide useful information, using the standards (RDF, SPARQL)
4. Include links to other URIs, so that they can discover more things.

(Berners-Lee, 2006)

The idea is to include links to concepts in the data whenever the data expresses what has been measured, where it has been measured, which method has been applied, and so on. Now it is easy to understand all those issues just by following links.

When this data is provided in RDF with a SPARQL endpoint, then you can search for any data that refers to a certain concept, otherwise you may maintain a concept based index of datasets just as you have it for documents. Finally terminology closes the gap between data and documents, and it opens new dimensions for data integration.

This application area is addressed by the new UFOPLAN 3712 12 100 Linked Environment Data project (Fock et al., 2012).

4. Summary and Conclusion

In the environmental information domain we have rather disparate levels of managed terminology. There is a quite stable general thesaurus both on the European and the national level (in Germany), but the underlying domains have no adequate representation.

We did not so much argue about motivation, we rather presented a set of methods to evolve and utilise managed domain terminology, which may motivate and support any existing effort in this direction. In case of questions, the new UFOPLAN 3712 12 100 Linked Environment Data project (Fock et al., 2012) may be able to help.
Bibliography


Twain (1880): Mark Twain, The Awful German Language. First published in the second volume of Twain's A Tramp Abroad, 1880, as appendix D.