LOD-ready Environmental Terminology with iQvoc

Thomas Bandholtz1, Joachim Fock2, Astrid Wolff3, Herbert Schentz4

Abstract

iQvoc is a new open source tool for maintaining and publishing domain terminologies on the Web following the “Linking Open Data” (LOD) pattern. This article gives a short overview of the underlying principles and introduces three environmental terminologies which have been deployed with iQvoc, each of them with its own characteristics.

1. Introduction

“LOD” is a buzz-word, short for “Linking Open Data”, driven by the more pragmatic fraction of the Semantic Web community. “LOD-ready” means providing environmental terminology in a way so that it can be referenced by linked datasets on the Web. In this article we will not talk about why we do so, but only how we do it. For a more comprehensive introduction into “Linked Environment Data” see (Rüther et al. 2010).

In continuous discussion with environment agencies all over the world6, the Federal Environment Agency, Germany, and innoQ Deutschland GmbH have developed an open source terminology tool named iQvoc7. iQvoc may be continuously developed based on contributions of more and more adopters, who add some functionality in change for using the tool for free.

So far, we have three running instances of iQvoc with different vocabularies, each of them having implemented some different “extras” on top of a shared, common standard.

2. General Concepts and Standards

This section clarifies the general concepts of terminology management and the related standards. Some application examples are presented in section 3. Before taking a closer look to these examples we should get a shared understanding of the underlying concepts and open standards.

2.1 Terms and Concepts

Any domain-specific (e.g. environmental) terminology makes an intellectual backbone of some shared information space. Any proper understanding and communication is based on sharing concepts and terms among the participants.

1 innoQ Deutschland GmbH, Halskestr. 17, 40880 Ratingen, Germany. thomas.bandholtz@innoq.com
2 Umweltbundesamt, FB I 1.5, Wörlitzer Platz 1, 06813 Dessau-Roßlau, Germany. joachim.fock@uba.de
3 Umweltbundesamt, FB IV 2.1, Wörlitzer Platz 1, 06813 Dessau-Roßlau, Germany. astrid.wolff@uba.de
4 Umweltbundesamt GmbH. Spittelauer Lände 5, 1090 Wien/Oesterreich herbert.schentz@umweltbundesamt.at
5 http://lod-cloud.net
6 Ecoterm: http://ecoterm.infointl.com;
7 https://github.com/innoq/iqvoc/wiki
Traditionally, concepts are denoted (and, hopefully, identified) by terms (or: labels), and at some place we may find a lookup (glossary, thesaurus, classification, taxonomy, ontology) of those terms and their intended meaning in our domain.

This traditional pattern has several drawbacks, based on ambiguities of certain lexicals: “bass” – a fish or a musical instrument? Michael Nagy gave a brilliant presentation about the fuzziness of the term “water use” (Nagy 2009), quoting eight different definitions by established authorities who disagree about significant patterns such as whether cooling water should be included or not.

Some of such drawbacks have been overcome by formalized thesauri which support cascading levels of granularity (concept hierarchy) and some diversity of terms that denote the same concept (synonyms, compositions, multilingualism). Here we may have one general concept of “water use” with eight narrower concepts such as “water use (OECD/Eurostat)”, “water use (UNSD/UNEP)”, and so on.

2.2 URI References

One drawback still persists: concepts are referenced by their labels (terms) which are finally represented by character strings. Such strings do not indicate any specific terminology source to be used as a lookup of their intended meaning. It certainly will make a difference looking up “water use” in the glossary of OECD or of UNEP. There is a ubiquitous ambiguity of any term representation. Character strings may be mistaken for terms, and terms may be mistaken for concepts.

Since some years, the Linking Open Data (LOD) pattern (Bizer et al. 2011) provides an exhaustive solution of this ambiguity problem: concepts are no longer referenced by terms (character strings) but by HTTP URIs (~ Web addresses) pointing directly to the definition of the concept in one dedicated terminology.

Certainly we do not want to replace terms by URI references in natural language documents in general. URI References are not made for human readers unless they can get resolved. While reading offline and printed, I cannot resolve URI references, so I need to have the term at hand. This is a common practice on the Web: the “term” appears as the “link title”, and the user sees this title, not the URI reference.

Using a named hyperlink whenever you mention a species or a pollutant may be tedious for the editor but illuminative for the data consumer: it takes just one click to tell a music instrument from a fish, or to figure out which concept of water consumption you are talking about.

2.3 Simple Knowledge System (SKOS) – a W3C Recommendation

Long before the Web, in 1974, we had an ISO standard (ISO 2788:1974) “Guidelines for the establishment and development of monolingual thesauri”. This standard has been revised in 1986 and is currently being replaced by ISO 25964 “Thesauri and interoperability with other vocabularies”.

As we are looking for a Semantic Web representation of thesauri, we favour a W3C Recommendation named Simple Knowledge Organization System (SKOS). SKOS is designed for thesauri, classifications, subject headings, taxonomies, and folksonomies, and, as we will see in section 3 and Fehler! Verweisquelle konnte nicht gefunden werden., we also use it as the basis of the gazetteer, the chronicle and the species service.

SKOS and ISO 25964 are very similar in structure, but only SKOS covers the Web representation in the context of Linked Data, so it comes closer to our needs. Anyway, it would be easy to convert a thesaurus from SKOS to the ISO XML-representation, lossless.

---

8 http://www.w3.org/TR/skos-reference/
Figure 1
The SKOS Core model of Concepts and Labels

Figure 1
SKOS XL Extension for Labels
SKOS has few classes (ConceptScheme, Concept, Collection and OrderedCollection), a set auf thesaurus-typic semantic relations plus a set of label and note properties. While semantic relations are bound to Concepts, label and note properties are devoted to the “Open World”: “anything” (every instance of owl:Thing) may use these properties, not only in a thesaurus.

The XL (“extension for labels”) provides means for expressing the complexity of nomenclature and lexical morphology, as we need it for or UMTHES (see section 3.1). In XL, labels are no longer simple character string values, but they are class instances with a property named literalForm. Here we put the character string (the term), and no we can have Label relations such as “abbreviated in”, “composedFrom”, and more. A closer discussion can be found in (Bandholtz 2009).

The Web representation of SKOS terminologies is based on the Resource Description Framework (RDF)\(^9\), which may appear in XML serialization, but also in N3, Turtle, or JSON, just to name some of those syntaxes. The main thing of RDF is that everything is expressed in triple form (subject, predicate, object), where subject and predicate must be URI references, and the object may as well be a literal value. Publishing something in RDF simply means making all these URI references resolve with a presentation of their associated triple statements. Now we can seamlessly navigate through this Web of Concepts and related data.

2.4 Implementation and Configuration

To make this scenario real, the terminology sources need to be published on the Web with unique and persistent URIs for each single Concept. This issue has been the main topic of the 5\(^{th}\) Ecoterm meeting in 2009 and of an EcoInformatics Web conference end of 2010\(^6\). Meanwhile some early adopters have provided a solution: GEMET and EUNIS (EEA), AGROVOC (FAO), and EARTH (CNR), each of them using different and more or less proprietary tools.

In order to simplify this procedure for those to come, the Federal Environment Agency in Germany (UBA, in the following “the Agency”) and innoQ Deutschland GmbH have developed an open source tool for the maintenance and publishing of terminologies in the LOD information space: iQvoc\(^10\).

iQvoc supports terminologies which can be represented in the SKOS Core model. iQvoc can further be configured and extended to support extensions of the SKOS model which are required by certain vocabularies. Basic examples of such extensions are described in the SKOS Primer (Isaac 2009). The following section gives three examples in detail.

The public iQvoc Website\(^10\) also includes a sandbox thesaurus (“The List of Hobbies”), where everyone can explore the editorial features of iQvoc and add his favorite hobbies.

3. Hosted Terminologies

At time when this is written iQvoc is present with three different environmental terminologies:

- German environmental thesaurus UMTHES\(^6\)
- Species Service of the German UBA
- Soil Thesaurus of the European GS Soil project.

So we have two thesauri of different complexity and one taxonomy. In the following there is a sub-chapter for each of them.

---

\(^9\) http://www.w3.org/RDF/
\(^10\) https://github.com/innoq/iqvoc/wiki
UMTHES® is a large formalized thesaurus based on the ISO 2788:1986. Migrating this to SKOS was quite challenging. SKOS comes closer to the new ISO 25964 model. But this is specific to the ISO/W3C models and not to a specific thesaurus. UMTHES® has three significant characteristics:

- German-centric multilingualism: preferred labels of each concept are given in German only, while English names are alternative labels. This means a restriction of the (intentionally symmetric) multilingualism of the SKOS model, but this is not in conflict with the standard.
- Exhaustive coverage of German lexical complexity including inflectional forms, spelling variants, and multiple possibilities of term composition, which was a serious pilot use case implementing the SKOS-XL extension in iQvoc.
- Poly-Hierarchy: Concepts may have multiple broader relations. Classifications usually have exactly one broader per Concept, and SKOS does not make any restriction in this question.

Konzepte

Bitte wählen Sie ein Konzept aus der unten aufgeführten Liste.

A B C D E F G H I J K L M N O P Q R

- A-Bewertung [Schall]
- Aal
- Aalen [Stadt]
- Aargau
- Aarhus-Konvention
- Aaskrähe [Corvus corone]
- Abbau
- Abbau [Bergbau]
- Abbau der Ozonschicht
- Abbau von natürlichen Ressourcen
- Abbaubarkeit
- Abbaubarkeit im Boden
- Abbaubarkeit im Wasser
- Abbauprodukt
- Abbauentemperatur
- Abbauvorgang

Figure 1
iQvoc alphabetical view showing UMTHES®
Today UMTHES® holds 13,170 Concepts with 13,170 preferred Labels plus 56,990 alternate Labels. Due to poly-hierarchy, there are more “broader” relations than Concepts: 17,722. There are 34 Concepts which have no broader relation at all, these are “Top Concepts”, and each of the other Concepts is a transitive child of at least one of those top Concepts. Besides the hierarchy, we also have 14,721 Concepts somehow “related” to each other in some a non-hierarchical way (without a closer typification). 3,283 Concepts point to remote Concepts from the GEneral Multilingual Environmental Thesaurus (GEMET)\textsuperscript{11} with a SKOS “mapping” relation.

These 13,170 plus 56,990 Label instances appear somehow peculiar with their multiple inflectional and may be composite lexical forms (Bandholtz 2009). Why are we maintaining this? This thesaurus is used with natural language processing, so it needs to know any possible representation of a concept by words in a written text. In a language like German with its “present” and “past” (<1996) orthographies and a highly inflected approach full of exceptions, Concept detection is completely different from the English experience: “Surely there is not another language that is so slipshod and systemless, and so slippery and elusive to the grasp”\textsuperscript{12}.

Below we give an English example which is comparatively simple, in RDF Turtle syntax\textsuperscript{13}:

```turtle
:4711 rdf:type skos:Concept;
  skosxl:prefLabel :wasteWater.
:wasteWater rdf:type skosxl:Label;
  skosxl:literalForm "waste water";
  sns:lexicalVariant "wastewater";
  sns:compoundFrom (:waste :water).
:waste rdf:type skosxl:Label;
  skosxl:literalForm "waste".
# only the noun, "wasted water" is NOT "waste water"!
:water rdf:type skosxl:Label;
  skosxl:literalForm "water";
  ext:inflectional "waters".
```

### 3.2 Species Service – a Reference of Test Organisms

The species service is a taxonomy of test organisms which have occurred in the Agency’s executive work. The Agency is responsible for the environmental risk assessment of plant protection products, biocides, pharmaceuticals and for chemicals which are regulated by the REACH Legislation.

As a basis for decisions for approval and evaluation of chemicals by the authorities, the applicants have to submit studies on bioaccumulation, biodegradation and toxicity of the substance. The submitted studies are evaluated for plausibility and validity, whereat certain criteria have to be met, for example, the test method and the test organism should be fully documented. The latter is ensured if a valid test species catalog is used for verification.

Such a comprehensive catalog was created through the merger and comparison of the test species catalogs from UBA and various national and international databases (e.g. IUCLID, the Chemicals Information

\textsuperscript{11}http://www.eionet.europa.eu/gemet
\textsuperscript{12}“The Awful German Language” by Mark Twain. http://www.gutenberg.org/files/119/119-h/p7.htm#Appendix_D
\textsuperscript{13}http://www.w3.org/TeamSubmission/turtle/
System of the EU). It contains about 6,000 valid test species from bacteria to invertebrates, fungi, algae, ferns up to vertebrates and vascular plants.

Figure 2
iQvoc hierarchical view showing an example of the species taxonomy

Quoting a domain expert:
"To date the catalog has got a new technological platform: The "Species Service" is now available as a web service based on the open source software iQvoc. The Species Service provides web navigation with alphabetical or hierarchical representation of the entire catalog. In the hierarchical representation, the highest hierarchical levels of biological classification; the phyla of animal and plant kingdom are shown and can be expanded up to the lowest hierarchical level of the species (possibly subspecies) in a tree structure. You can click on any level of the tree to get a taxon that includes the scientific name of the organism, its parent and child(ren) in the biological classification. In addition, common names, synonyms and other information, such as the habitat, are shown. Each taxon has its own web address in accordance with the standards of "Darwin Core". Furthermore the Species Service has a search function for (parts of) Scientific, German or English names. For registered users editing functions are available to modify and to create a new taxon. A service interface is available to connect the Species Service directly with information systems such as ICS (Information System on Chemical Safety) (Menger et al. 2011), which is used in the process of risk assessment by the UBA."

As said, the species catalogue is formalized as a taxonomy. Typically, species taxonomies use scientific names as preferred labels and different types of alternative labels in multiple languages, such as trivial names or outdated scientific names. Scientific names may be understood as Latin names like "Puma con-
color (Linnaeus, 1771) sec. Brown”, but not every Latin name of a taxon is the scientific name, so we need a dedicated sub-property of the preferred Label.

Different from UMTHES®, there is a mono-hierarchy with well-defined levels (taxon ranks), e.g. Kingdom, Phylum, Class, Order, Family, Genus, Species. There is no general standard of these ranks and their representation. The Taxonomy Database Working Group (TDWG)¹⁴ is developing a format called “Darwin Core”, but this still consists of elements without a finalized structure. We extended the SKOS model (and iQvoc) to support scientific names and taxon ranks from Darwin Core in a close-to-SKOS way.

In a next step, mapping relations should be established, for example with the “Pan-European Species directories Infrastructure” (EU-Nomen)¹⁵, or with the “European Nature Information System” (EUNIS)¹⁶.

### 3.3 A European Soil Thesaurus

The GS-Soil project¹⁷ is working on “Assessment and strategic development of INSPIRE compliant Geo-data-Services for European Soil Data” and has started to develop a soil-specific thesaurus (“SoilThes”)¹⁸. SoilThes is “is intended to support finding and understanding soil and soil metadata information”. It will be available in at least 10 languages and integrate Concepts from multiple sources, such as GEMET, INSPIRE Themes, ISO11074 soil vocabulary, or AGROVOC. In this case the challenge is supporting a distributed team of domain experts and translators in a long-term cooperative effort, and the idea of a “federation of terminologies” (see section 0) is brought up.

SoilThes will be built on the SKOS Core model, with English as the “main language”. Because of lack of time and resources, preferred Labels will always be from one predefined language, while any other language will appear in alternate Labels only. A lesson learnt from GEMET was that each language has its specific set of concepts, which GEMET does not express. For example, “wood” and “forest” are semantically not equivalent to German “Wald” and “Forst”. The GEMET definition of “wood” reads: “A dense growth of trees more extensive than a grove and smaller than a forest.”¹⁹ In German, a “Forst” is not necessarily larger than a “Wald”, but it is a “Wald” with some economic utilization. For such reasons, using preferred Labels in each language for the same Concept may be semantically misleading. SoilsThes and UMTHES® both have decided on Concepts of a single language, with the closest names from other languages as alternate Labels only. For the GSSoil community it is clear, that the usability of SoilThes beyond the runtime of the project depends on the possibility to maintain the controlled vocabulary.

iQvoc needed some optimization with regard to handle multilingualism in distributed editorial teams. There will be editors in the role of translators who want to see the English Labels and definitions while editing in their own language. Maybe they also want to see existing translations in a third language. So we decided on an individual multi-language selection. The editor now may read English and French while he is writing Spanish. Further languages are completely hidden from the user interface in this state. Multi-language selection can as well be used by any public user who is just browsing the terminology.

---

¹⁴ http://www.tdwg.org/
¹⁵ http://www.eu-nomen.eu/portal/index.php
¹⁶ http://eunis.eea.europa.eu
¹⁷ http://www.gssoil.eu/
¹⁸ https://secure.umweltbundesamt.at/soil/en
¹⁹ http://www.eionet.europa.eu/gemet/concept?langcode=en&cp=9349&ns=1
4. On-going Work: Services and Federation

Since 2003 the Agency provides its “Semantic Network Service” (SNS)\(^{20}\), technically the predecessor of iQvoc. SNS includes UMTHES, a gazetteer and a chronicle with cross-links among each other, and it provides Web Services such as similar search term proposals or automatic classification of documents.

The gazetteer will extend SKOS with elements of the Geonames Ontology\(^{21}\) (gn:Feature) and a spatial intersection property. Here we have a hierarchy for administrative units only, but a river is neither “broad-er” nor “narrower” than a natural landscape, to give just one example.

The chronicle will extend SKOS with elements of the Linked Events Ontology\(^{22}\) for temporal properties. Until now we have no hierarchy of events in the data, but this may be used in the future, for example to indicate a workshop is part of a conference.

Currently we are re-developing the SNS services from scratch, based on an 8 years’ experience. The first use case to be supported by iQvoc (instead of “old school” SNS) will be assisted keyword assignment for Catalogue datasets in INSPIRE.

One architectural difference between SNS and iQvoc is in how they manage multiple interlinked terminologies: one instance of SNS is hosting three terminologies, while each instance of iQvoc will host exactly one. Whatever a single SNS instance is applying to multiple terminologies must be implemented over a communication protocol shared by each of the federated iQvoc instance.

As this federation will be based on open interfaces, the participating terminologies are not necessarily hosted in iQvoc instances. Any terminology service may participate if it implements these interfaces. This architecture also provides a modular scalability as more terminologies may join over time without any modification of the federation service. Features of the federation service should be:

- A registry of the participants (VoID\(^{23}\) plus some possible extension);
- A human readable presentation of this registry with links to the root nodes of each terminology’s Web representation.
- Administrative services such as
  - a workbench for terminology alignment and cross-linking terminology and data.
- End user services such as
  - searching all terminologies simultaneously,
  - distributed auto-indexing documents and data based on selected terminologies,
  - an integrated, concept based index of Web documents and data

5. Summary and Conclusion

For today, we can draw two conclusions: iQvoc is ready for use, and the open source model has proved beneficial for all participants. Each of the terminology providers has sponsored some specific extension or optimisation, and he can profit by the work which has been sponsored by the others. Finally, iQvoc has emerged as a valuable tool which is available to any terminology provider for free.

Three reference terminologies have been made LOD-ready in a technical sense, simply by importing them into an iQvoc instance on the Web. However, till today, they are literally LOD-ready, but still there is no data linked to them. There is some on-going work in this direction in the Linked Environment Data

\(^{20}\) http://www.semantic-network.de
\(^{21}\) http://www.geonames.org/ontology/documentation.html
\(^{22}\) http://linkedevents.org/ontology
\(^{23}\) http://www.w3.org/TR/void/
initiative (Rüther et al. 2010, 2011), but the process is very slow and stepmotherly funded by the authorities. Amazing it is, as Linking Open Data is the “golden bullet” for establishing a Shared Environmental Information System (SEIS)\(^\text{24}\) where “information is provided once and shared with others for many purposes”. IQvoc is a free building block in such an information architecture.

**Bibliography**


\(^{24}\) http://ec.europa.eu/environment/seis/