Linked Environment Data: SCOVO-fying the Environment Specimen Bank

Thomas Bandholtz
innoQ Deutschland GmbH
Halskestr. 17
40880 Ratingen, Germany
+49 178 4049387
thomas.bandholtz@innoq.com

Till Schulte-Coerne
innoQ Deutschland GmbH
Halskestr. 17
40880 Ratingen, Germany
+49 1515 7132845
till.schulte-coerne@innoq.com

Maria Rüther
Federal Environment Agency
Corrensplatz 1
14195 Berlin
+49 30 89031503
maria.ruether@uba.de

ABSTRACT *(of submission)*

In this paper, we discuss the proposed RDF schema and external linkage of the German Environmental Specimen Bank (ESB). The schema is based on SCOVO and SKOS with some powerful extensions. The submission raises several strategic SCOVO design issues that are currently subject of discussion in the Linked Data community.

At the time of submission, the ESB has been published with less SCOVO extensions (like in example 3) in a test version.

Categories and Subject Descriptors

E.1 [**Data Structures**]: Distributed Data Structures

H.5 **[Information Interfaces and Presentation]**

J.3 **[Life and Medical Sciences]:** Health

General Terms

Design, Standardization.

Keywords

Triplification, Vocabularies, Linked Governement Data, Statistical Data, Environmental Data, Content Negotiation, SPARQL.

1. INTRODUCTION

Currently several projects at the German Federal Environment Agency (UBA) begin with the design and implementation of a public data network that is technologically based on Linked Data. The first ones will be the Environmental Specimen Bank (ESB) and the Semantic Network Service (SNS); the inclusion of the Dioxin Database and the Joint Substance Data Pool of the German Federal Government and the German Federal States (GSBL) is currently under discussion. The undertaking is an international cooperation jointly with the Ecoterm Initiative and the European Environment Agency (EEA), and it is envisioned to include the partners of the International Environmental Specimen Bank Group (IESB).

These projects and partners provide the key instruments in the field of environmental observation that enable the long-term analysis of substance exposure of humans and the environment.

2. ESB Schema Based on SCOVO

The ESB reports the accumulation of pollutants/substances in test subjects at specific places with respect to time.

Basically, this can be expressed with the Statistical Core Vocabulary (SCOVO) [1].

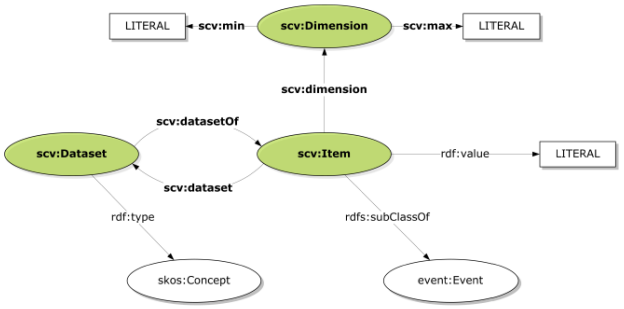

Figure 1. Original SCOVO Model (from [1]).

SCOVO is modelled around scv:Item, which links to several scv:Dimension individuals, and a scv:Dataset. It provides a simple rdf:value Literal.

In the referred SCOVO publication [1], the flight on-time arrival example has three dimensions: *OnTime*, *Airport*, and *TimePeriod*. The value is given as a simple number like in the following example:

```
ex:myItem a scv:Item ;
    rdf:value 74 ;
    scv:dataset ex:ontimeFlights ;
    scv:dimension ex:Q12006 ;
    scv:dimension ex:onTimeArrival ;
    scv:dimension ex:AtlantaHartsfield .
```

Example 1. Original SCOVO example (from [1]).

However, the simple SCOVO model needs some tailoring and extension in order to represent the complexity of the ESB model. One aspect is about representing hierarchical dimension parameters by linking SCOVO and SKOS, the other is about tailoring subclasses of scv:Dimension according to the ESB model.

2.1 Competency Questions

Before going into the details, it is a good idea to consider some recommendations about the design of an "exemplary ontology"[1], one of which is that "explicitly stated requirements (especially competency questions) were specified and used to guide the design process". Such questions might be:

1. What types of dimensions are supported by the ESB?
2. What are the valid combinations of dimension types in a data item?
3. Are there different datasets representing different combinations of dimension types?
4. What dimension individuals are supported for each dimension type?
5. What does a given dimension individual exactly mean?
6. Given an individual item, what are the summary statistic and the unit of measurement for the provided value?
7. What is the time line of a given set of dimension individuals?
8. What is the spatial distribution of a given set of dimension individuals?
9. More specifically: what was the impact of the latest extreme Elbe flood on the chemical exposure of limnetic samples?
10. More specifically: does the mercury exposure of test persons increase significantly when they use amalgam fillers?

The simple model of the original SCOVO example can answer several of these questions with respect to its simple data. The ESB however adds complexity in several aspects which we will discuss one by one in the following.

2.2 Specializing scv:Dimension

Compared to the SCOVO example, ESB has four major dimensions:

- *SpecimenType* – in most cases, a specific part of a test subject, e.g. the "bream liver".
- *Analyte* – what has been analyzed with respect to the specimen type. Analytes are split into *Substances* and *Biometrics* (see section 2.4 below).
- *Sampling area* – at which place has the sample been taken.
- *TimeReference* – when has the sample been taken. This is given as a full year only.

So far, this is structurally similar. The adapted example would look as follows:

```
esb:specimen a scv:Item ;
    rdf:value 2.1672 ;
    scv:dataset esb:chemicalExposure ;
    scv:dimension esb:breamLiver ;
    scv:dimension esb:Copper ;
    scv:dimension esb:lowerRhine.Bimmen ;
    scv:dimension esb:year2008 .
```

Example 2. Adapted ESB example.

2.2.1 Data Type Considerations

Still something is missing: what does rdf:value 2.1672 mean? In the original SCOVO example, the value 74 means "74 *percent* of on-time arrivals …" but this has not been stated explicitly in the data, so the value might be easily misunderstood as meaning a *count* of on-time arrivals.

Tauberer [2] tries to solve this for the 2000 U.S. Census by providing the value in a string such as `"122283145776 m^2"`, for square meter which he calls "a strange half-number/half-unit literal value" himself. Obviously, this might also be expressed as a typed literal[2] such as

```
"122283145776"^^ex:squareMeter.
```

But what to do if the publisher wants to specify that the value should be interpreted as an xsd:double the same time?

In the ESB, we have several complex units of measurement (such as "microgram per litre wet weight"), *and* we have several summary statistics (such as "geometric mean" or "standard deviation") for most values. The first approach of the ESB project was to extend the rdf:value attribute, thereby allowing to provide a complex data structure as each item's value. However, this would result in reduced flexibility and proprietary processing requirements. The preferred solution is a further extension of the dimension set by adding two scv:Dimension subclasses:

- *UnitOfMeasument* and
- *SummaryStatistic*.

From a data warehouse point of view, those two might not be understood as dimensions but as metadata of different scv:DataSet individuals. In our opinion, this design would result in another loss of flexibility in the SCOVO representation.

2.2.2 Item Filtering vs. Dataset Hierarchies

Furthermore, the ESB supports anamnesis filters (such as "gender", "smoker", "amalgam fillers", etc.) for human test subjects. The overall value of a human sampling collective can be split in subgroups by such filters.

Tauberer [2] has a similar structure in the 2000 U.S. Census, and his design (when translated to SCOVO) is a hierarchy of datasets, where each dataset is bound to one of these "filters". Alternatively, this could be modelled as another hierarchy level of the dimension (see 3 below). However, we find this too restrictive: if you have a dataset for smokers and another one for females, how then to query the values of female smokers?

For this reason, we added another dimension.

- *AnamnesisFilter*.

In summary, a complete example of an ESB item might look like:

```
esb:specimen a scv:Item ;
    rdf:value 0.1800 ;
    scv:dataset esb:chemicalExposure ;
    scv:dimension esb:24hSamplingUrine ;
    scv:dimension esb:Greifswald ;
```

[1] http://ontologydesignpatterns.org/wiki/Odp:WhatIsAnExemplaryOntology

[2] http://www.w3.org/TR/rdf-concepts/#dfn-typed-literal

```
    scv:dimension esb:Mercury ;
    scv:dimension esb:year2008 ;
    scv:dimension esb:geometricMean ;
    scv:dimension esb:mgPerLitreWetWeight ;
    scv:dimension esb:female ;
    scv:dimension esb:smoker .
```

Example 3. Extended ESB example.

Example 3 shows concise information but somehow difficult to read by itself, as we have eight scv:dimension properties without a local indication of the respective dimension type. Following the example from [1], this information is not contained in the scv:Item itself. Somewhere else in the esb: namespace you will find statements like:

```
esb:SamplingArea
    rdfs:subClassOf scovo:Dimension .

esb:Greifswald a esb:SamplingArea .

esb:Analyte
    rdfs:subClassOf scovo:Dimension .

esb:Substance
    rdfs:subClassOf scovo:Analyte .

esb:Mercury a esb:Substance .

esb:Biometric
    rdfs:subClassOf scovo:Analyte .

esb:Weight a esb:Biometric .
```

Example 4. Dimension subclasses and individuals.

In order to understand the meaning of each scv:dimension property, one has to look up the type definition of the provided individual in the object role. The SCOVO publication [1] contains a SPARQL example for this use case.

2.3 Sub-Properties of scv:dimension

It might be more convenient to setup sub-properties of scv:dimension such as esb:samplingArea, and give them a range of a respective subclass of esb:Dimension, in this case esb:SamplingArea.

```
esb:samplingArea
    rdfs:subPropertyOf scv:dimension ;
    rdfs:range esb:SamplingArea .

esb:substance
    rdfs:subPropertyOf scv:dimension ;
    rdfs:range esb:Substance .
```

Example 5. scv:dimension subProperty examples.

This would result in a more readable item serialization as given in the following example:

```
esb:specimen a scv:Item ;
    rdf:value 0.1800 ;
    scv:dataset esb:chemicalExposure ;
    esb:specimenType esb:24hSamplingUrine ;
    esb:samplingArea esb:Greifswald ;
    esb:substance esb:Mercury ;
    esb:timeReference esb:year2008 ;
    esb:summaryStat esb:geometricMean ;
    esb:uom esb:mgPerLitreWetWeight ;
    esb:anamnesisFilter esb:female.
    esb:anamnesisFilter esb:smoker.
```

Example 6. More expressive ESB item example.

So far this is more expressive but the choice of dimension types for an item is still arbitrary. In the next step we will restrict this using subclasses of scv:Item.

2.4 ESB Subclasses of scv:Item

The ESB supports to different types of datasets:

- *chemicalExposure*
- *biometrics.*

Chemical exposure describes the amount of some substance which has been measured in some human or environmental test subjects, while biometrics describe specific aspects of test subjects (age, gender, weight, etc.), so there is a different set of analytes in use in the biometrics dataset.

Complementing the ESB subProperties of scv:dimension, one might define subclasses of scv:Item and use them in the domain role of these properties. This could bind a certain item type to an intended set of dimension types. Consequently we would use the respective dimension subclasses in the range of these properties.

```
esb:ChemicalExposureItem
    rdfs:subClassOf scv:Item .

esb:substance
    rdfs:domain esb:ChemicalExposureItem .

esb:BiometricItem
    rdfs:subClassOf scv:Item .

esb:biometric
    rdfs:domain esb:BiometricItem .
```

Example 7. Sublasses of svc:Item with dedicated properties.

This idea is somehow seductive from an object-oriented point of view but Linked Data is bound to the open world assumption. In this context, "valid" combinations of dimension subclasses are simply those which are used in the published scv:Item individuals. Such provided combinations can be explored by dedicated SPARQL queries against the published data.

When the ESB subclasses and subProperties are used, OWL semantics will infer that each individual used in the object role is a member of the class which has been specified in the range statement.

Given the schema so far, competency question 2 (valid combinations of dimension types) can be answered by a schema look-up (all subProperties of scv:dimension with a given subclass of scv:Item as their domain).

2.5 Adding Cardinality Restrictions

In the ESB, where most dimension types are mandatory for one of the item types, we have only one optional dimension: esb:anamnesisFilter. The following example describes that each ChemicalExposureItem must have exactly 1 esb:substance property.

```
esb:ChemicalExposureItem
    rdfs:subClassOf [
        rdf:type owl:Restriction
        owl:onProperty esb:substance ;
        owl:cardinality 1
    ] .
```

Example 8. Cardinality restriction .

We would express a corresponding restriction for each dimension type except from the anamnesisFilter, as it may appear even multiple (see Example 6), or it may not appear. So far, the generic serialization from Example 3 is still valid, as it refers to scv:Item, not to one of the more restrictive subclasses.

2.6 scv:Item, Event, and Time

The original SCOVO schema[3] models scv:Item as subclass of "Event". This refers to the Event Ontology[4] which has been published in 2007. In the context of Linked Data, this was superseded by the Linked Events Ontology[5], and so SCOVO the scv:Item superclass may be changed accordingly.

In both cases, we find this subclassing misleading: both Event classes have temporal and spatial properties on their own, but SCOVO ignores these properties and uses the dimension model instead. The Event superclass is not used in theses examples at all, why then do we need it?

The discussion in [1] gives a rather philosophical rationale: "An event is then defined in this ontology as the way by which cognitive agents classify arbitrary time/space regions. Our Item concept is subsuming this Event concept - a statistical item is a particular classification of a time/space region". Obviously there is some semantic similarity between the two concepts, but given the formal schemas of each, scv:Item inherits several properties from Event which are semantically in conflict with its own dimension properties. On those terms, we recommend to drop this subclassing.

On the other hand, there is an issue about modelling the *Time Reference* as a dimension individual. If we follow SCOVO and specify a time reference like in example 9 below, then the scv:min and scv:max properties are suggesting a kind of temporal sort order.

```
esb:year2002 a esb:TimeReference ;
    scv:min "2002-01-01"^^xsd:date ;
    scv:max "2002-12-31"^^xsd:date ;
```

Example 9. Time reference with min and max.

But how to express a timeline of items starting in 2002? There is no way to express something like "later than esb:year2002", as esb:year2002 is a RDF resource and not a typed literal. A possible solution would be to provide the time reference not in form of dimension individuals but as a datatype property of scv:Item. In fact we cannot imagine any statistical dataset without a time reference, so this could be a convenient built in-property of the generic scv:Item. The Event subclassing might be helpful from this point of view. However, none of the two Event ontologies considers the concepts of "later" and "earlier", and neither does the "W3C Time Ontology in OWL"[6].

[3] http://sw.joanneum.at/scovo/schema.html

[4] http://motools.sourceforge.net/event/event.html#Event

[5] http://linkedevents.org/ontology/

[6] http://www.w3.org/TR/owl-time/

2.7 scv:Dataset Consideration

The original definition of scv:Dataset is rather vague: "a dataset, representing the container of some data, such as a table holding some data in its cells".

From a data warehouse point of view, scv:DataSet corresponds to the "fact" table. In this table, a specific set of dimension individuals makes up the primary key, and this key is bound to the corresponding measured value. Each scv:Item individual corresponds to one row (not to one cell) in a fact table.

In the discussion about using SCOVO in the Vocabulary of Interlinked Datasets (VOID), there is one remarkable statement by Richard Cyganiak:

"In SCOVO, scovo:Items are grouped into scovo:Datasets, and there seems to be an implicit assumption that all items in such a dataset share the same dimensions."[7]

Strictly speaking and referring to our model of the ESB, Richard does not mean "share the same dimensions" but "share the same set of scv:dimension sub-properties".

This exactly is what we have modelled in the ESB subclasses of scv:Item in section 2.4. So we do not need the scv:Dataset construct at all. Subclassing scv:Item gives everything what scv:Dataset wants to express, and, moreover, it provides the described means to restrict a dataset to the respective dimension's sub-properties.

3. ESB Dimensions, SCOVO, and SKOS

SCOVO offers several "hooks" [1] for linking statistical data with domain ontologies, two of them are "subclassing the SCOVO dimension class", and the "built-in support" for skos:Concept. We will discuss both approaches in the following sections.

3.1 SKOS for Dimension Hierarchies

Figure 1, taken from the original publication of SCOVO [1], shows skos:Concept as the "type" of scv:Dataset, "in order to allow hooking into a categorisation scheme" [1]. First of all, both skos:Concept and scv:Dataset are classes, so we would prefer to make scv:Dataset a subclass of skos:Concept, not a skos:Concept individual.

What's more important: we think that SKOS is helpful for structuring dimensions rather than datasets. In most statistical data collections (just like the ESB), the dimension individuals, such as sampling areas, are organized in a hierarchy, and there is an individual hierarchy for each dimension type. The ESB Web application provides a hierarchy of "profile" pages for each dimension, so we already have a linked tree of information resources. This tree can simply be rendered in SKOS to provide some simple but valid domain ontology.

```
esb:dimensions a skos:ConceptScheme ;
    skos:hasTopConcept esb:specimenType.

esb:specimenType a skos:Concept ;
    skos:narrower esb:limneticSample .

esb:limneticSample a skos:Concept ;
    skos:broader esb:specimenType ;
    skos:narrower esb:bream .
```

[7] http://code.google.com/p/void-impl/issues/detail?id=18

```
esb:bream a skos:Concept ;
    skos:broader esb:limneticSample ;
    skos:narrower esb:breamLiver .
```

Example 10. skos:ConceptScheme for ESB dimensions.

In the SCOVO scheme, scv:dimension has no explicit range, so we can simply use skos:Concept individuals in the object role of scv:dimension property statements like in the previous examples (see esb:breamLiver in Example 2).

However, we have modelled subProperties of scv:Dimension and assigned them a range of a specific subclass of scv:Dimension in section 2.3. Thus, by using any of the skos:Concept individuals in the object role of, say esb:substance, we can infer that this is also an esb:Substance individual.

Now we want some explicit class assertions on that in order to clarify the intention of the data source independently from any open world usage.

3.2 Bringing SKOS and SCOVO Together

So far, this results in subclasses of scv:Dimension (such as esb:SpecimenType, esb:Substance, etc.) on one side, and a skos:ConceptScheme tree with the corresponding top concepts on the other. Which of the skos:Concept individuals belongs to what scv:Dimension subclass is only inferred by its usage in some scv:Item instance.

If we want to make this more explicit, we could use a double class assertion (skos:Concept *and* the scv:Dimension subclass). Alternatively, we make scv:Dimension a subclass of skos:Concept, and so the SKOS tree (Example 10) and the dimension subclassing (section 2.2) integrates to the following.

```
scv:Dimension
    rdfs:subClassOf skos:Concept .

esb: SpecimenType
    rdfs:subClassOf scv:Dimension .

esb:dimensions a skos:ConceptScheme ;
    skos:hasTopConcept esb:specimenType.

esb:specimenType a esb: SpecimenType ;
    skos:narrower esb:limneticSample .

esb: limneticSample a esb: SpecimenType ;
    skos:broader esb:specimenType ;
    skos:narrower esb:bream .

esb: bream a esb: SpecimenType ;
    skos:broader esb:limneticSample ;
    skos:narrower esb:breamLiver .
```

Example 11. scv:Dimension as subclass of skos:Concept.

This pattern has also been proposed by Jeni Tennison [7]. If you are setting up your RDF dimension tree from scratch, this avoids double class assertions for each individual.

If you already have a SKOS tree which you want to utilize as dimension tree, you may want to express: all individuals which are in a sub-tree with given top concept as its root shall be individuals of a corresponding subclass of scv:Dimension. This can be formalized if you use skos:broaderTransitive instead of skos:broader. Using broader with the transitive flavour, all individuals in the sub-tree will have this top concept as their skos:broaderTransitive by inference. In this case you can avoid explicit double class assertions for each individual by the following OWL restriction:

```
esb:SpecimenType
    rdfs:equivalentClass [
        rdf:type owl:Restriction
        owl:onProperty
skos:broaderTransitive ;
        owl:hasValue esb:specimenType
    ] .
```

Example 12. esb:SpecimenType class assertion for skos:Concept individuals by restriction on skos:broaderTransitive .

4. ESB External Linkage

Hausenblas et al. [1] recommend linking the dimension individuals to some domain ontology which has been published in the Web, or to the more general DBpedia using owl:sameAs.

In the case of the ESB, we have a built-in domain ontology which is published together with the statistic data; however, this does not make a big difference: we are linking the ESB domain ontology to several external vocabularies anyway.

This happens in an international context of cooperating governmental authorities, such as the eTerminology Workshop [5] or the Ecoterm Group[8] with members from many European countries and the US. These authorities have started setting up a trusted network of domain ontologies [4] from different environmental facets and multiple languages. Such ontologies will be published in RDF (namely SKOS) and described in the Vocabulary of Interlinked Data (VOID). Some of them have already gone live or will be going live in the next months.

The interlinking of the ESB will focus on these vocabularies and further environmental measurement data to be published in the Web as well. Governmental authorities are quite reserved against non-governmental vocabularies such as DBpedia or Geonames. Publications of the agencies usually are subject to legal obligations and thus sensitive about the provenance of the sources. Currently there is no decision about such linkage.

In the following, we will give a short overview of the intended linkage and the structure and state of each target.

4.1 Semantic Network Service of the Federal Environment Agency

Semantic Network Service (SNS)[9] is maintained by the agency since 2003. SNS includes a thesaurus (UMTHES), a gazetteer and a chronicle with occasional interlinkage among each other. All three are currently available in the XML Topic Maps[10] format. A first draft of an RDF vocabulary for SNS has been presented in 2006, but until today only the thesaurus has been migrated into a SKOS-XL representation.

[8] http://ecoterm.infointl.com

[9] http://www.semantic-network.de

[10] http://isotopicmaps.org

The gazetteer contains the ESB sampling areas (among others). The RDF schema may extend the Geonames ontology[11] by a domain specific type system and some properties for spatial intersections between individuals which are not organized in a hierarchy (e.g. river crosses city).

The chronicle will be interlinked with the ESB time references. For example, the Elbe flood in 2002 had a considerable impact on the exposure of limnetic samples in this area in the following year. Such relations can be expressed by linking the respective statistical items (in this case not the dimensions) to this event.

The chronicle can be published in the Linked Event Ontology[12] with a domain specific type system, and a somehow simplified pattern of describing the time references.

4.2 More Vocabularies from the Ecoterm Space

The Ecoterm group members represent more than 20 different vocabularies but we will focus on only three for the beginning.

The European reference vocabulary since years is the GEneral Multilingual Environmental Thesaurus (GEMET)[13], maintained by the European Environment Agency (EEA). GEMET has been one of the first SKOS use cases in 2004 and is still available in this serialization. Since last year it is also published using the Linked Data technical patterns.

UMTHES is already linked with GEMET, so we do not need any direct linkage between ESB and GEMET. GEMET is much smaller than UMTHES (which has been one of its sources) but it is available in 29 languages.

The second vocabulary from the EEA is the EUNIS biodiversity database[14], with a focus on species. EUNIS has been published in RDF early this year, using several properties from the Darwin Core vocabulary[15].

The third example is the Environmental Applications Reference Thesaurus (EARTh)[16] from Italy, which has been published in SKOS and linked with EUNIS as well.

4.3 Environmental Information Systems

All those partner vocabularies mentioned above are reference vocabularies, not data about the state of the environment. ESB provides such data, and we will link it to more data. Currently, there is some environmental data published within the Data-gov projects of the US[17] and UK[18].

Within the realm of the Federal Environment Agency in Germany, there are two such systems which may follow in the near future:

- The Dioxin Database[19] is built on the same model as the ESB and currently accessible through Web Services. Once we have established the ESB in Linked Data, the same patterns can be adapted to this database.
- The Joint Substance Data Pool of the German Federal Government and the German Federal States (GSBL)[20] contains detailed dossiers about substances, some of which occur as analytes in the ESB. The ESB substances have already been linked to these dossiers, but the GSBL is based on a completely different database model, and it will take some extra effort to get this published in RDF.

Finally, there is ongoing work to establish an International Environmental Specimen Bank Group (IESB)[21] with partners from Scandinavian countries, Canada, Japan, South Korea, France, UK, and US.

4.4 owl:sameAs Consideration

Most Linked Data contributors (including [1]) tend to use owl:sameAs for cross-references. Following the OWL reference, “Such an owl:sameAs statement indicates that two URI references actually refer to the same thing: the individuals have the same "identity"."[22]

This has two implications: (1) make sure about exact semantic identity, and (2) a reasoner will merge both individuals, including their class assignments.

(1) There are cases where two things appear to be the same but are not. In the ESB, for example, two different species, the bream and the deer, have a skos:narrower, which might be named “liver”. Someone might state that both are the same and link them to some anatomy vocabulary talking about the abstract concept of “liver”. Most certainly, however, a deer liver is not the same as a bream liver. SKOS provides a more subtle set of mapping relations[23] skos:closeMatch, skos:exactMatch, skos:broadMatch, skos:narrowMatch, and skos:relatedMatch. Using these, one might express that esb:breamLiver and esb:deerLiver both have a skos:broader relation to anatomy:liver, which would come much closer to the domain knowledge than owl:sameAs.

(2) This brings up the merging issue. Using owl:sameAs, the esb:breamLiver (a skos:Concept) and anatomy:liver (some other class) would be treated as just one individual which now belongs to both classes. In the respective schemas, there may be collisions between the two class definitions. In the case of the SKOS mapping relations, domain and range of these properties would make both individuals instances of skos:Concept. Such implications may be ignored as long as linked data gets simply browsed, but they will result in semantic collisions as soon as a reasoner or some more specific agent is traversing the Web of Data.

[11] http://www.geonames.org/ontology

[12] http://linkedevents.org/ontology

[13] http://www.eionet.europa.eu/gemet

[14] http://eunis.eea.europa.eu/

[15] http://rs.tdwg.org/dwc/

[16] http://uta.iia.cnr.it/earth_eng.htm

[17] http://data-gov.tw.rpi.edu/wiki/The_Data-gov_Wiki

[18] http://data.gov.uk/

[19] http://www.pop-dioxindb.de/

[20] http://www.gsbl.de/eng_home.htm

[21] http://www.inter-esb.org/

[22] http://www.w3.org/TR/owl-ref/#sameAs-def

[23] http://www.w3.org/TR/skos-reference/#mapping

The ESB provides scientific data that can give input for further research and integration with different sources, so we will link it carefully.

5. Exploring the ESB

Feigenbaum [5] describes his approach of exploring the Eurostat LOD contribution (which has not been published in SCOVO) by querying the graph with SPARQL.

This is something that will work with any published RDF data that supports a SPARQL endpoint. ESB will support multiple approaches to explore its structure and data.

5.1 Starting at the ESB Homepage

One starting point is the homepage of the ESB Web application at www.umweltprobenbank.de. This application provides various information resources about the ESB, a tree of interlinked, human readable dimension profiles, and a comfortable HTML/JQuery database interface. Results are displayed in tables or diagrams which can be exported in CSV and Excel formats. We are planning to make most of this accessible in RDF in several ways:

1. Each dimension profile will support content negotiation as described in [6];
2. For human visitors with standard browsers and no linked data plug-in there will be direct links to the respective RDF representation;
3. The query dialog will provide the URI of the SPARQL access point and display the SPARQL version of each query.

Complementing this, there will be a Linked Environment Data introduction page with a description of the project and links to the documented RDF schema, VOID metadata, and to the SPARQL endpoint. The VOID metadata will be linked to the Ecoterm project.

5.2 Coming from an Interlinked Vocabulary

Our intention is to have bi-directional links between the ESB and the partner systems listed in chapter 4. Whenever an agent explores one of these partner systems, she will detect outgoing links pointing to some related ESB dimensions or items. From here the agent can explore the linkage of this resource as described in section 5.3.

Bi-directional links may not be realistic in some cases, as this requires some extra effort from the partner agencies. We can provide a concise list of the cross-references to minimize this effort but we cannot incorporate such links into the partner system ourselves.

If bi-directional links cannot be established, the upcoming Ecoterm platform may help: “a simple umbrella Web page would be created to aid in the management, promotion, and access to this network of linked data” [4]. The first step will be linking to the VOID metadata of all member vocabularies from this umbrella page. If each of these VOID datasets links back to the umbrella page at least, this would offer an indirect path to the interlinkage: from any ex1:resource to the ex1: namespace to the ex1: VOID to the Ecoterm platform to the ex2 VOID dataset which describes the linkage from ex2 to ex1. Ecoterm could also host lists of cross-references. Currently, this is still a vague idea which needs some clarification and agreement in detail.

5.3 Exploring the Neighbourhood of a Single ESB resource

There are various scenarios which lead to a single ESB resource as the starting point. The previous section described bi-directional linkage but you may also find such a link in a human readable document or an e-mail.

In most cases, this link will lead to an ESB dimension or item individual.

In case of a scv:Dimension individual, ESB offers different names in German and English, a short description in the domain context, and links to further information resources and related data. As we modelled scv:Dimension as a subclass of skos:Concept, you may traverse the dimension tree level by level. Using SPARQL, one may retrieve all statistical items that refer to this individual (which will be quite many), or inspect the existing combinations with different dimension subclasses first.

Having an item as the starting point, you may simply move on to explore each of the connected dimension individuals for a better understanding of the meaning of this item. Using SPARQL, you may drop one of the dimensions to get a list of related items, e.g. a time line, or a comparison of different sampling areas at the same time.

These are only some simple examples. Being aware of the ESB schema, SPARQL gives you nearly unlimited options.

6. Summary

(Preliminary, will be updated in the camera-ready version).

This is the first time a German Authority is publishing measurement data as Linked Data. We tried to learn from the SCOVO discussion and developed a semantically “strong” schema, including a SKOS concept scheme:

- The SCOVO model can be applied with extensions.
- ESB needs an extended set of scv:Dimension subclasses.
- ESB uses explicit statements about units of measurement and summary statistics.
- There is an issue about the temporal sort order and comparison (“earlier”, “later”) of time references.
- Subclasses of scv:Item are used to specify different datasets and bind them to mandatory sub-properties of scv:dimension
- ESB integrates its domain vocabulary by making scv:Dimension subclass of skos:Concept.

Several projects of different environmental authorities plan to establish detailed cross-references between vocabularies, datasets, and other information resources:

- Each dimension individual links to various information resources for extended understanding in different media types.
- Those information resources should link back to the dimension individual.
- ESB will be carefully linked to multiple reference vocabularies from the Ecoterm space. Bi-directional links are envisioned.

- ESB will be linked to further data collections of environmental authorities.
- ESB (Germany) initiates a close linkage between different international ESBs.

7. REFERENCES

[1] Hausenblas, Michael; Halb, Wolfgang; Raimond, Yves; Feigenbaum, Lee; Ayers, Danny. 2009. SCOVO: Using Statistics on the Web of Data. ESWC 2009. http://sw-app.org/pub/eswc09-inuse-scovo.pdf.

[2] Tauberer, Joshua. 2007. The 2000 U.S. Census: 1 Billion RDF Triples. rdf:about. http://www.rdfabout.com/demo/census/

[3] Bandholtz, Thomas et al.. 2009. Summary of W4 eEnvironment Terminology. e-envi2009. Integrating Environmental Knowledge in Europe. March 25-27, 2009 Prague. http://www.e-envi2009.org/SummaryTerminologyW4.pdf

[4] Ecoterm Group. 2009. Report on the outcome of the Ecoterm V Workshop. UN Food and Agriculture Organization, Rome, Italy on 5-6 October 2009 http://eea.eionet.europa.eu/Public/irc/envirowindows/jad/library?l=/ecoinformatics_indicator/ecoterm_5-6102009

[5] Feigenbaum, Lee. 2008. Modeling Statistics in RDF - A Survey and Discussion. TechnicaLee Speaking. http://www.thefigtrees.net/lee/blog/2008/03/modeling_statistics_in_rdf_a_s.html

[6] Bizer, Chris; Cyganiak, Richard; Heath, Tom. 2007. How to Publish Linked Data on the Web. http://www4.wiwiss.fu-berlin.de/bizer/pub/LinkedDataTutorial

[7] Tennison, Jeni. 2009. Expressing Statistics with RDF. Jeni's Musings. http://www.jenitennison.com/blog/node/132