Linked Environment Data

Maria Rüther, Joachim Fock, Joachim Hübener
Umweltbundesamt
Wörlitzer Platz 1, D-06813 Dessau-Roßlau
maria.ruether@uba.de
joachim.fock@uba.de
joachim.huebener@uba.de
Thomas Bandholtz, Till Schulte-Coerne
innoQ Deutschlang GmbH
Halskestr. 17, D-40880 Ratingen
thomas.bandholtz@innoq.com
till.schulte-coerne@innoq.com

Abstract

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[1]. 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 still under discussion. The undertaking is part of an international cooperation in the *Ecoterm Initiative*, and it is envisioned to include the partners of the International Environmental Specimen Bank Group (IESB)[2].
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.

1. Linked Data and Environmental Informatics

Since the 1990's, the linking of environmental data and technical vocabularies is one of the UBA's main goals which has been pursued since several project generations (UMPLIS, UDK, GEIN, SNS, PortalU). All previous efforts, however, have two common drawbacks:

- Up to now, only data containers (databases, information systems, complex Web pages) have been linked together – and not individual data records.
- There is no common access to a shared data structure, so that each cross-reference ends at the doors of the referenced data store, or, in the best case, at a Web page describing the access.

Linked Data is different: a network of individual data elements linked together for direct access and navigation on the Web. The linking mechanism is based on Web addresses (HTTP URIs) for each data element and on the universal data model of the Resource Description Framework (RDF).

---

[1] http://linkeddata.org/
[2] http://www.inter-esb.org/

In 2006, Tim Berners-Lee formulated four "Linked Data Principles"[3]:

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.

Though some of this wording may be discussable, the Semantic Web Community has published Billions of RDF triples and links in the following years. Starting from any point in this part of the Web, one can easily "discover more things" click by click. The "Linked Data Cloud" (Figure 1) has become a huge playground for the exploration of this technology.

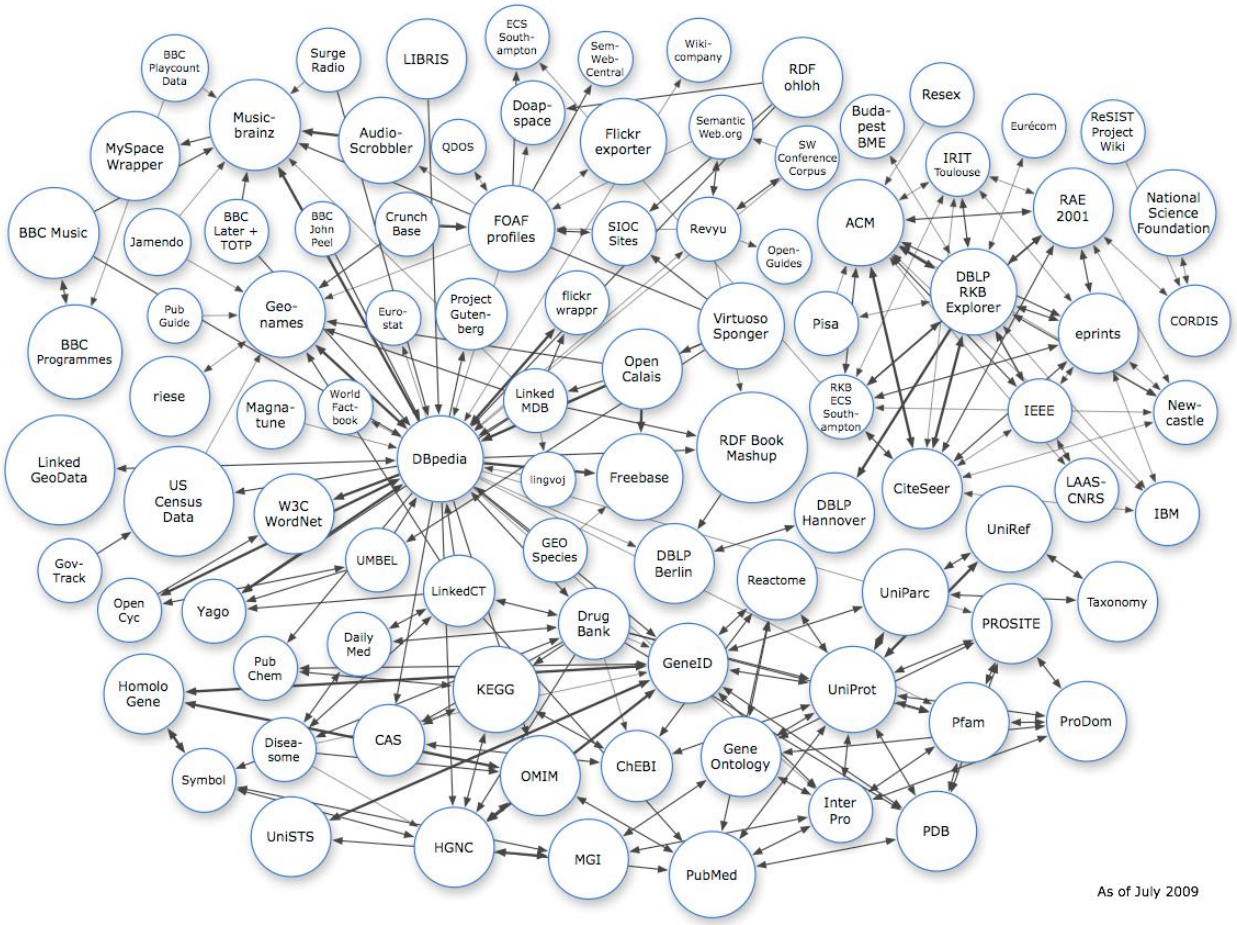

Figure 1 Linked Data Cloud (from[4])

All this may be seen as yet another example of (more or less academic) community enthusiasm around the Linked Data representation of Wikipedia, DBpedia.

[3] Tim Berners-Lee, 2006-07-27 http://www.w3.org/DesignIssues/LinkedData.html

[4] http://richard.cyganiak.de/2007/10/lod/

However, there are also more serious, scientific efforts. The most elaborated example is the Linking Open Drug Data[5] (LODD) sub-cloud in the EHealth community.

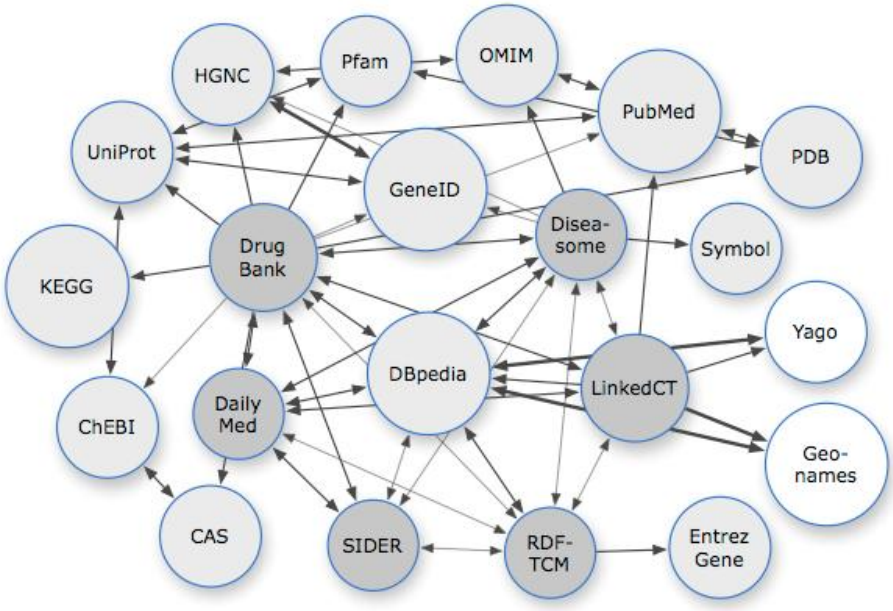

Figure 2 Linking Open Drug Data (LODD)

2. Linked Environment Data

The LODD example gave birth to the idea of linking environmental data in an international context of co-operating governmental authorities.

This idea was discussed at Workshop V (Ecoterm 2009) of the Ecoterm Group[6] with members from many European countries and the US. Ecoterm fosters "*a federated approach to accessing terminology and knowledge organization systems in the area of the environment that would allow them to be accessed, interchanged, and used in traditional indexing and search approaches, as well as semantic web applications. The idea is to share the content of these rich resources in such a way that duplication of effort can be avoided and interchange and integration of various structured and unstructured data can be enhanced. The approach should allow the vocabularies to be linked over time, as appropriate, and for resources to be linked to these vocabularies.*" (Ecoterm 2009)

The focus is clearly on setting up reference vocabularies. Participating authorities have started setting up a trusted network of domain ontologies from different environmental facets and multiple languages. Such ontologies are going to be published in RDF (namely SKOS) and described in the *Vocabulary of Interlinked Data* (VOID)[7]. Some of them have already gone live or will be going live in the next months.

The commitment to linking environmental data to these vocabularies is rather vague: the "SKOS files would also be published as linked data", but someone else has to publish the observation data and link it to the reference vocabulary.

[5] http://esw.w3.org/HCLSIG/LODD
[6] http://ecoterm.infointl.com
[7] http://vocab.deri.ie/void/guide

Figure 3 shows such plans from Germany. Blue bubbles stand for SKOS vocabularies in the Ecoterm context, white squares for information systems holding observation data.

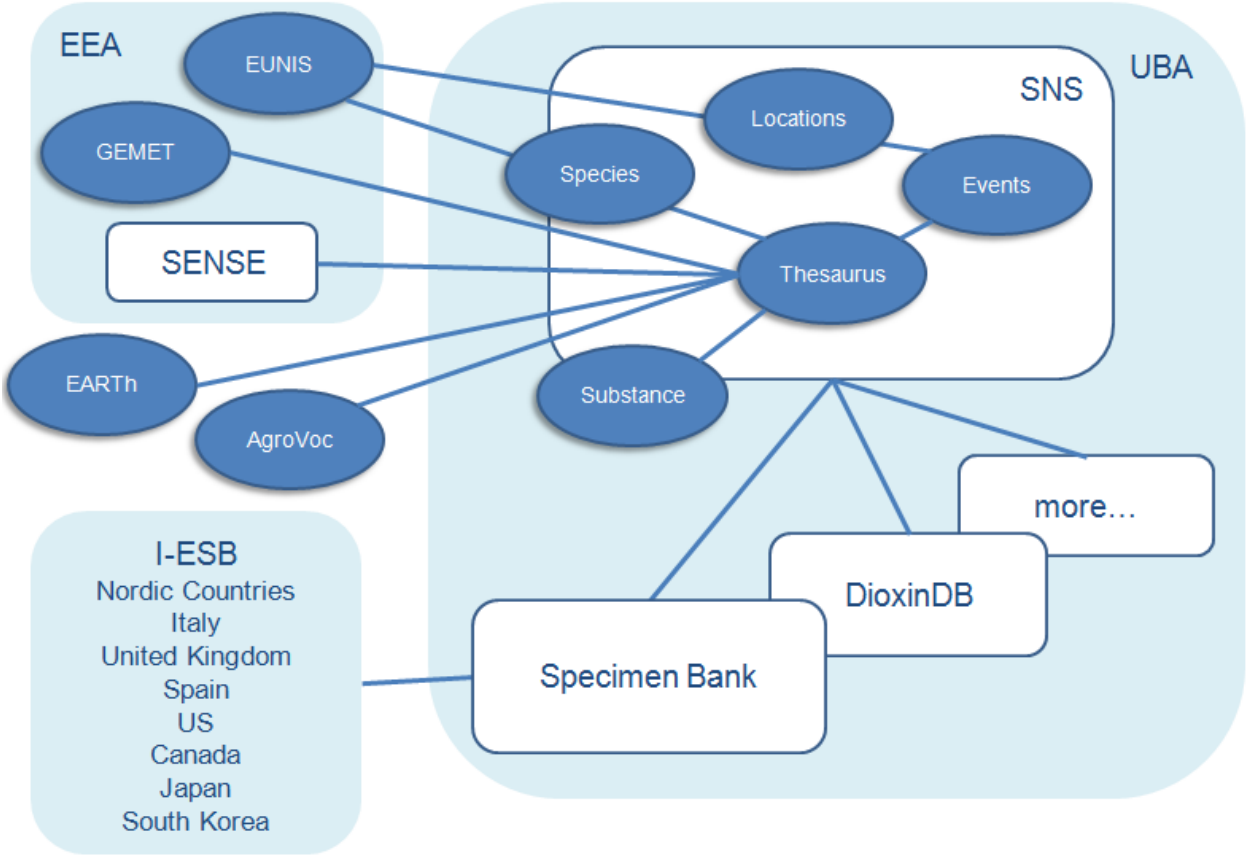

Figure 3 Linked Environment Data (Vision@de 2009)

Semantic Network Service (SNS)[8] is maintained by the UBA 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[9] 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 (sees more details about the RDF models of SNS in section 3).
Figure 3 shows links from SNS to several European vocabularies (top left).
The European reference vocabulary since years is the GEneral Multilingual Environmental Thesaurus (GEMET)[10], 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.

[8] http://www.semantic-network.de
[9] http://isotopicmaps.org
[10] http://www.eionet.europa.eu/gemet

The second vocabulary from the EEA is the EUNIS biodiversity database[11], with a focus on species. EUNIS has been published in RDF early this year, using several properties from the Darwin Core vocabulary[12].
The third example is the Environmental Applications Reference Thesaurus (EARTh)[13] from Italy, which has been published in SKOS and linked with EUNIS as well.
The bottom of Figure 3 shows some exemplary observation data which we plan to publish as linked data in this context. The German ESB is the starting point in this case, and we will try to motivate international partners (I-ESB) to join this Linked Data cloud.

The ESB reports the accumulation of pollutants/substances in defined samples at specific places with respect to time but is not itself the specialist that can exhaustively describe these reference elements. Hence, the data has to be linked to specialized information about each of these parameters. For substances, for example, the links could point to the corresponding substance information in the GSBL, for species (as test subjects) to the EUNIS[14], for places to the Geo Thesaurus of the SNS, for time references to the Environment Chronicle (SNS). The Environmental Thesaurus (UMTHES) forms a layer on top of it, which in turn is linked to the international GEMET.
Each data record of the ESB can be directly linked to the pieces of information of these specialized services. Ideally, the specialist information links back to the data records, thereby enabling bi-directional navigation.
Additionally to all the previously mentioned information systems, there are numerous specialists that are not provided by authorities, e.g. Chemical Entities of Biological Interest (ChEBI)[15], or GeoNames[16]. The question, whether these are to be linked as well, is a political one: The technological prerequisites are fulfilled.

3. RDF Representation

In order to put the linking mechanism to work and being able to directly access a given reference, a RDF data representation for all participating systems needs to be created. It is based on HTTP URIs (Web addresses) and a generic data model that has triples (subject/predicate/object) as its sole constituent. Subject and predicate are always encoded as HTTP URIs, the object can be an URI as well, or a literal (e.g. a number or a character string). For examples, please refer to the participants' models in the following sections.
This approach forms the basis for describing and applying individual models (RDF Schema or „vocabulary") that are broadly comparable to object-relational models but can be semantically richer. Numerous RDF vocabularies have already been established. These vocabularies can and should be used, combined, and extended whenever possible and needed.
In the following we use an ESB data example (Figure 4) with a striking peak in 2004.

[11] http://eunis.eea.europa.eu/
[12] http://rs.tdwg.org/dwc/
[13] http://uta.iia.cnr.it/earth_eng.htm
[14] http://eunis.eea.europa.eu/species.jsp
[15] http://www.ebi.ac.uk/chebi/
[16] http://www.geonames.org/

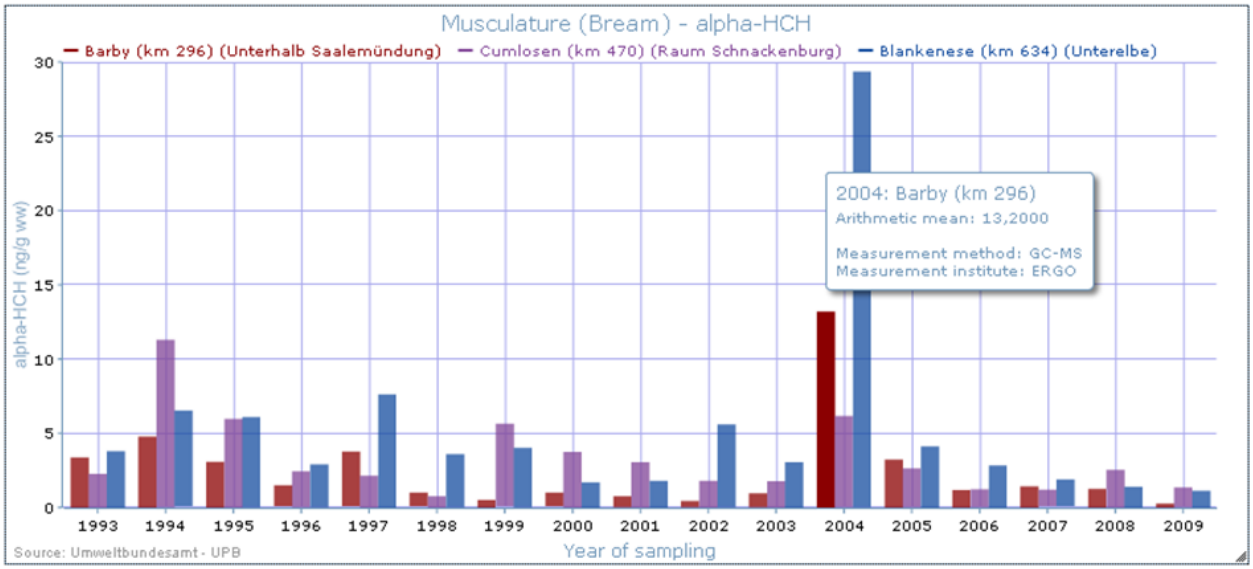

Figure 4 ESB data example

It is assumed that during earthworks and dike reconstructions following the Elbe flood of 2002, high quantities of alpha-HCH and beta-HCH were released from soils and dumping grounds around Bitterfeld and entered the river Elbe via the river Mulde. The increasing Mulde-contamination is reflected in elevated levels of alpha-HCH and beta-HCH in bream since 2003 with peak concentrations in 2004 resp. 2005 in both, the river Mulde and river Elbe.
The RDF examples in the following sub-chapters will demonstrate how this peak gets annotated by links between ESB and SNS.

3.1 RDF Model for the Environment Specimen Bank (and DioxinDB)

The ESB's data model (similar to that of DioxinDB) can be represented based on the Statistical Core Vocabulary (scovo)[17]. SCOVO has been developed in order to publish statistical data from Eurostat and other sources in the Linked Data cloud [Hausenblas et al., 2009].
We have proposed some extensions and specializations in order to represent the domain-specific dimensions (specimen type, analytes, sampling area), that is, the classifications of the ESB profiles (Rüther 2010).
In parallel, the Publishing Statistical Data group[18] has started an effort of integrating SCOVO with the *Statistical Data and Metadata eXchange* (sdmx.org) guidelines, which leads to a different approach. Currently the upcoming extended SCOVO version is an open issue. The example in Figure 5 shows the data of the 2004 peak following our own proposal (may be subject of change till final production).

Looking at these RDF examples one must be aware this is a machine readable format behind the human-friendly Web pages (more on this in section 4). In this format we make heavy use of abstract IDs (UUIDs or just numbers) which is perfect for machines but not for human reading. So we added some comments in the code (starting with an "#") in order to make it more human readable.

[17] http://sw.joanneum.at/scovo/schema.html
[18] http://groups.google.com/group/publishing-statistical-data

The examples are given in Turtle syntax[19], and – just like in XML – we are using namespace prefixes such as rdf:, esb:, etc. as abbreviations of full URIs in most cases.

Important links are marked in red.

```
esb:t47098a_10220d1bc3e_4ee1 a scv:Item ;
   rdf:value 13.2000;
   scv:dataset esb:chemicalExposure ;
   esb:specimenType esbd:10006 ;
   esb:samplingArea esbd:10137 :      # Blankenese (Unterelbe)
   esb:substance esbd:10053           # alpha-HCH ;
   esb:timeReference esbd:year2004 ;
   esb:summaryStat esbd:arithmeticMean ;
   esb:uom esbd:nggww ;
   sns:relatedEvent sns:t1d97d0d_102035cd5d4_-3384. # Elbe flood
esbd:10006 a esbd:SpecimenType ;
   skos:prefLabel "Brassenmuskulatur"@de ;
   skos:prefLabel "bream musculature"@en ;
   skos:broader esb:10037 .
esbd:10037 a esbd:SpecimenType ;
   skos:prefLabel "Brassen"@de ;
   dwct:scientificName "Abramis brama" ;
   owl:sameAs <http://eunis.eea.europa.eu/species/9986> .
```

Figure 5 ESB data snippet (the „peak“) in RDF

The first code block ends with a sns:relatedEvent statement which point to the event which has caused the peak. The code snippet ends with an equivalence statement (owl:sameAs) pointing to the EUNIS representation of Abramis brama (the bream).

3.2 RDF Model of the Environmental Chronicle

The chronicle has not yet been published in RDF so far. Figure 6 shows a snippet from a draft. The final RDF model should consider the Linked Events Ontology[20], which itself is an extension of the „An Ontology of Time for the Semantic Web“[21].

In the bottom of this code example you see links to the location and the descriptor of this event.

[19] http://www.w3.org/TeamSubmission/turtle/

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

[21] http://www.w3.org/2006/time

```
sns:t1d97d0d_102035cd5d4_-3384 a sns:Event;
   skos:prefLabel „Extreme Flooding on Elbe River"@en;
   dct:description „During the first 13 days of August 2002
   extraordinary heavy rains fell in parts of Central Europe,
   causing disastrous flooding along the rivers Elbe and Danube
   and severe damage totalling an estimated 30 000 million US $.
   450 000 people were evacuated and more than 100 lost their
   lives. In the first 5 days of August heavy thunderstorms
   developing in warm moist air affected northern Germany. On 1
   August, some stations reported the highest 24-hour rainfall on
   record (Cuxhaven 63,6 mm), and regionally, the precipitation
   totals of 2 days exceeded the monthly normals by 50 %."@en;
   sns:location sns:FLUSS5;
   sns:descriptor umt:_00028876;
   …
```

Figure 6 SNS chronicle snippet of the Elbe flooding

3.3 RDF Model of the UMTHES Thesaurus

The RDF model of the UMTHES has already been implemented [Bandholtz 2009]. It is based on the Simple Knowledge Organisation System (SKOS)[22] vocabulary, making heavy use of the "extension for labels" (SKOS-XL)

```
umt:_00028876 a skos:Concept;
   skosxl:prefLabel :Hochwasser;
   skosxl:altLabel :Flusshochwasser, :Flut--Hochwasser,
   :Flutereignis, :Flutkatastrophe, :Fruehjahrshochwasser,
   :HochwasserEinesFlusses, :Hochwasserereignis,
   :Hochwasserganglinie, :Hochwasserganglinienvorhersage,
   :Hochwasserkatastrophe, :Hochwasserrisiko, :Oderhochwasser,
   :Winterhochwasser, :flood, :floodLevel, :floodWater,
   :highTideWater, :highwater;
   skos:broader :_00650524, :_00028887, :_00027345;
   skos:narrower :_00012775, :_00012789, :_00012791, :_00651102,
        :_00012793;
   skos:related :_00029767;
   skos:exactMatch <http://www.eionet.europa.eu/gemet/concept/3298>;
   sns:descriptorOfEvent sns:t1d97d0d_102035cd5d4_-3384 ;
   [...] .
```

Figure 7 UMTHES RDF snippet of the flood concept

In the bottom there is a link to GEMET (skos:exactMatch) and a back-link to the event shown in Figure 6.

3.4 RDF Model of the Gazetteer

The gazetteer contains the ESB sampling areas (among others). The RDF schema may extend the Geonames ontology[23] 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).

[22] http://www.w3.org/2004/02/skos/

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

.

```
sns:FLUSS5 a sns:Location;
   sns:locationType sns:river ;
   skosxl:prefLabel :Elbe ;
   skosxl:altLabel :ElbeRiver ;
   owl:sameAs <http://sws.geonames.org/2931271/> ;          # ???
   sns:spatialIntersection :BIOSPHAERE1, :NATURPARK16,
        :NATURRAUM881, :GEMEINDE1207005032,
        :WASSEREINZUGSGEBIET537, GEBIRGE21453, [...] ;
   sns:locationOfEvent :t1d97d0d_102035cd5d4_-3384,
        :t392814_101fe3b006f_956 ;
   [...] .
```

Figure 8 Gazetteer RDF snippet of the river Elbe

Here we find an equivalence link to Geonames (which might be arguable), several spatial intersections (non-hierarchical!), and a back-link to the event shown in Figure 6.

4. Technological Architecture

The technological aspects of Linked Data publication is described in detail in [Bizer 2007]. However, speaking in terms of efficiency, it is not advisable that each of the participating information systems implements these mechanisms independently. The German Federal Environment Agency implements a dedicated Linked Data server. At the time this gets written, there is a test environment based on Virtuoso Universal Server[24] (Open Source edition), but also BigOWLIM[25] may be considered. Acting as a common proxy, the Linked Data server would de-reference all URIs, forward to the HTML representation of each system if needed, and moreover provide a SPARQL endpoint.

This server would also take care about content negotiation (Figure 9), a most important facet of Web technology.

In this pattern, we have one URI for the abstract concept, in this example: the bream. However, this URI cannot resolve in anything, as we are not able to download “the bream”. We just need it to indicate when we are talking about the bream in general (“Use URIs as names for things”).
Whenever an agent sends a resolving request to such an URI it will provide the preferred content type (mime-type). In case of a Web browser, this will always be HTML, so you will never see any RDF representation in a standard Web browser when resolving the concept URI (though there are several plug-ins …). In case of specific RDF aware data agents, they can make use of several RDF syntax mime-types. The server inspects the preferred mime-type of the agent and sends a redirect to the corresponding specific document URI. While we cannot download “the bream”, we can download documents about the bream. And we can as well talk about a certain document representation of the bream. That is why these documents have distinct (document) URIs.

[24] http://virtuoso.openlinksw.com/
[25] http://www.ontotext.com/owlim/big/index.html

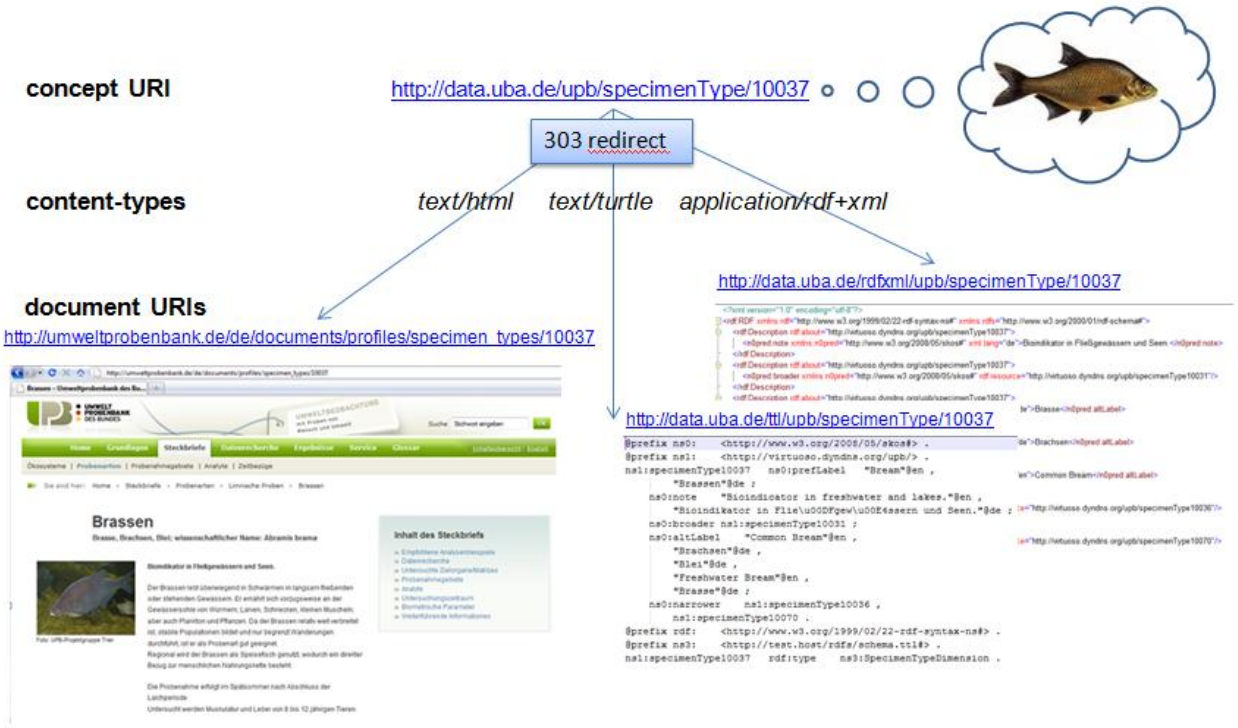

Figure 9 Content Negotiation

In this Linked Data Server scenario, each participating system would only need to render its own data records in the corresponding RDF vocabulary and post changes to the Linked Data server.
What is more, this architecture would allow for the implementation of further visualization services, e.g. like the ones already in evaluation by the Data-gov project of the U.S. government[26].

5. Literature

Bandholtz, Thomas: Expressing Lexical Complexity in SKOS-XL. Ecoterm Rom 2009. http://eea.eionet.europa.eu/Public/irc/envirowindows/jad/library?l=/ecoinformatics_indicator/ecoterm_5-6102009&vm=detailed&sb=Title

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

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

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

Rüther,Maria; Bandholtz, Thomas; Schulte-Coerne, Till; SCOVO-fying the Environment Specimen Bank. Draft Feb 2010, http://www.w3.org/egov/wiki/images/8/85/Isem2010-bandholtz.pdf

[26] http://data-gov.tw.rpi.edu/wiki/Demo:_Castnet_Ozone_Map