Linked Environment Data

Maria Rüther, Joachim Fock, Joachim Hübener
Umweltbundesamt
Wörlitzer Platz 1, D-06813 Dessau-Roßlau
maria.ruether@uba.d
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 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)[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

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 was no data structure with common access, so that each cross-reference ended at the doors of the linked data store, or, in the best case, at a Web page describing the access.

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

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

2. Linking the Environmental Specimen Bank

The ESB reports the accumulation of pollutants/substances in test subjects 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[3], 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)[4], or GeoNames[5]. The question, whether these are to be linked as well, is a political one: The technological prerequisites are fulfilled.

3. RDF Models

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.

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

The ESB's data model (similarly to that of DioxinDB) can be represented with the Statistical Core Vocabulary (scovo)[6] [Hausenblas et al., 2009]. Some extensions will be necessary in order to represent the domain-specific dimensions (specimen type, analytes, sampling area), that is, the classifications of the ESB profiles.

3.2 RDF Model of the UMTHES

The RDF model of the UMTHES has already been implemented [Bandholtz 2009]. It is an extension of the Simple Knowledge Organisation System (SKOS)[7] vocabulary.

[3] http://eunis.eea.europa.eu/species.jsp

[4] http://www.ebi.ac.uk/chebi/

[5] http://www.geonames.org/

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

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

3.3 RDF Model of the Environmental Chronicle

The Environmental Chronicle is part of the Semantic Network Service (SNS)[8]. Until now, the SNS has used the XML Topic Maps format, which is not compatible with RDF. An independent design of an RDF vocabulary for SNS has already been presented in 2006. From today's perspective, it would be appropriate to leverage new technological developments and check the suitability of the Linked Events Ontology[9], which itself is an extension of the „An Ontology of Time for the Semantic Web“[10].

3.4 RDF Model of the Geo Thesaurus

The Geo Thesaurus is part of SNS as well. Since its design in 2006, the GeoNames Ontology[11] has gained in importance, most probably due to its intelligent use through Geonames.org. For the description of coordinates, WGS84 Geo Positioning[12] has been widely accepted since 2003.

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 authors strongly advocate that the German Federal Environment Agency implements a dedicated Linked Data server. Acting as a common proxy, it would dereference all URIs, forward to the HTML representation of each system if needed, and moreover provide a SPARQL endpoint.

In this scenario, each participating system would only need to output 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[13].

5. Literature

Bandholtz, Thomas: Expressing Lexical Complexity in SKOS-XL. Ecoterm Rom 2009.

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/

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

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

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

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

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

[12] http://www.w3.org/2003/01/geo/wgs84_pos

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