Linked Environment Data

Getting Things Connected

Thomas Bandholtz[1], Joachim Fock[2]

[1]innoQ Deutschland GmbH, Monheim am Rhein, Germany
thomas.bandholtz@innoq.com
[2]Federal Environment Agency (UBA), Dessau-Roßlau, Germany
joachim.fock@uba.de

Abstract. After three years of discussion and early prototypes, the Federal Environment Agency (UBA), Germany, now has launched a two-year research & development project on Linked Environment Data (LED) with innoQ Deutschland GmbH as a contractor. This project will set up a core cloud of environment data with a well-elaborated domain terminology as its semantic backbone. Data will be taken from the "Environmental Specimen Bank", the "German Metadata Portal on Soil" and further databases such as the "Joint Substance Data Pool of the German Federal Government and the German Federal States" as well as the environmental library and research databases. The infrastructure will support a sustainable process of keeping the data permanently up-to-date, and there will be a dynamic and intuitive user interface. All the work will be fully Semantic Web compliant, based on vocabularies such as SKOS, SCOVO or Data Cubes, and Dublin Core.

Keywords. Environmental protection, domain terminology, observation data, linking open data.

1 Introduction

Networking among comprehensive observation data and domain terminology has been a basic concern of the UBA since the 1990s with various project generations (named UMPLIS, UDK, GEIN, SNS and PortalU). All these implementations so far have two common weaknesses:

- The linkage established by these systems has connected data containers (data bases, information systems, complex Web pages) but not individual data records.
- There was no shared data structure to be accessed for exploitation, so that every link ended up so to say in front of the door of the referenced database, at best on a Web page describing the respective data access.

Linked Data, however, stands for linking individual data records that can be easily dereferenced. Tim Berners-Lee has summarized the four principles already in 2006 [1]:

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."

In 2009 he added a "5 star rating" to make this more clear and to acknowledge the Linking *Open* Data movement:

* "Available on the web (whatever format) but with an open license, to be Open Data
** Available as machine-readable structured data (e.g. excel instead of image scan of a table)
*** as (2) plus non-proprietary format (e.g. CSV instead of excel)
**** All the above plus, Use open standards from W3C (RDF and SPARQL) to identify things, so that people can point at your stuff
***** All the above, plus: Link your data to other people's data to provide context"

Here we see that Linked Data has been envisioned without an explicit demand of "openness" in mind, and actually Linked Data can be perfectly applied within closed communities as well.

The environmental authorities in Europe have a strong tradition of publishing open data which has been expressed by the Aarhus Convention [2] in 1998 and Directive 2003/4/EC on public access to environmental information [3] in 2003. So 1 and in parts 2 and 3 star data has been provided by these authorities since years. While there certainly is some remaining discussion about legal limitations of this openness, the real input is the "Linked" aspect in this domain, which has been described more in-depth by Tom Heath and Chris Bizer in 2010 [4].

The vision of Linked Environment Data came up at the eTerminology workshop [5] at the e-Envi conference in Prague in March 2009 and was elaborated during the 5th Ecoterm meeting [6] in Rome in October of the same year.

In 2010 the European Environment Agency made the General Environmental Multilingual Thesaurus (GEMET)[1] and the European Nature Information System (EUNIS)[2] available as Linked Open Data, followed by the Environmental Applications Reference Thesaurus (EARTh)[3] provided by Istituto Inquinamento Atmosferico in Italy. In December there was a 2-day Ecoinformatics International Webinar on

[1] http://www.eionet.europa.eu/gemet/
[2] http://eunis.eea.europa.eu
[3] http://uta.iia.cnr.it/earth_eng.htm

Linked Open Data[4]. LED was also discussed by the W3C eGovernement Interest Group[5] and topic several conference contributions.

In 2011, the German "Umwelt-Thesaurus" UMTHES[6] has been published as Linked Data as well, and a (strictly non-open) species taxonomy in the context of substances approval. There was an early (open) Linked Data test-bed of the German Environmental Specimen Bank (ESB)[7] which was not deployed into production. The yearly EnviroInfo[8] conference hosted a full day session on „Linked Open Data, Semantic Search and Interoperability", and there will be a follow-up in 2012: "Linked Environment Data – Getting Things Connected".

However, these early implementations have been rather scattered and have dominant focus on domain terminology, not so much observation data. In a „Use Case Crosslinking Environment Data and the Library"[9] you can read about the German contributions: "The most prominent obstacle is the lack of a dedicated funding for this initiative. There are some projects of the participating systems that draw up some of their budget for pieces of the puzzle, but there is no overall plan of the agency so far."

This use case drafts a scenario where observation and library data get cross-linked among each other and with the domain terminology which has been seized by the Linked Environment Data research & development project (UFOPLAN 3712 12 100) finally launched by the German agency by the time this is written.

2 Strategic Issues of the LED Project

2.1 Master Plan and Project Portfolio

By end of 2012 there will be a master plan, inter-coordinated with all stakeholders, which provides a strategic foundation beyond the borders of the two-year project. There will be prioritised work packages, some of which may be implemented in 2012 as well.

The overall portfolio will be highly dependent on how far the corresponding projects can work on their interfaces themselves or have to delegate this to the LED project. Currently we cannot make certain assumptions.

In any case we aim for a - more or less comprehensive – pilot system (or pilot cloud) which makes the aspired "added information value through interlinked data" a real experience. Moreover there must be a demonstration of how the standardised RDF interfaces and the LED workbench simplify the integration of further data.

---

[4] http://projects.eionet.europa.eu/ecoinformatics/library/ecoinformatics_indicator/meeting_6-7122010

[5] http://www.w3.org/egov/wiki/Linked_Environment_Data

[6] http://data.uba.de

[7] http://umweltprobenbank.de

[8] http://www.ec-gis.org/Workshops/EnviroInfo2011

[9] http://www.w3.org/2005/Incubator/lld/wiki/Use_Case_Crosslinking_Environment_Data_and_the_Library

2.2 Project Infrastructure

During the first month we will decide on the project infrastructure together with the computer centre of the agency. It will consist of:

- Production system with man/machine interface (content negotiation)
- Triple store
- Registry based on the vocabulary of interlinked data sets (VoID)[10]
- Cross database data-recall client
- (geo-)graphic visualisation services
- Workbench with tools enabling RDF interfaces and data-linking

One special part of this infrastructure is iQvoc[11], an open source terminology management tool that we have developed jointly over the last two years.

All this is glued together by a careful selection and extension of standardised RDF vocabularies such as VoID, SKOS[12], SCOVO[13] or Data Cubes[14] which are "understood" and interpreted by the machine.

The registry will know which participant uses which standard und can even describe local extensions, so that code extensions are not necessary. Of course such extensions have a limited freedom, which needs to be defined and communicated.

2.3 Integration and Extension of Existing Approaches

The existing LED prototypes of the agency have to be aligned with the LED master plan. They all include native methods for RDF data rendering and can synchronise with a triple store incrementally. However, these methods have been developed and need to be revisited, refactored, and extended. The same applies to the RDF formats and the linkage.

Environment Specimen Bank (ESB)

The Environmental Specimen Bank records the accumulation of (harmful) substances in defined samples at certain locations and times. However the ESB itself is not responsible for the comprehensive description of all relevant elements, so specialized information should be referenced instead. For substances such data is provided by GSBL, for species there is EUNIS, for locations and times SNS's geo thesaurus and environmental chronicle, respectively. The environmental thesaurus (UMTHES) provides an overarching envelope which is in turn linked with the international GEMET.

[10] http://www.w3.org/TR/void/
[11] https://github.com/innoq/iqvoc
[12] http://www.w3.org/2004/02/skos/
[13] http://vocab.deri.ie/scovo
[14] http://publishing-statistical-data.googlecode.com/svn/trunk/specs/src/main/html/cube.html

In the early test-bed the ESB data model was represented in SCOVO, but today we consider the Data Cubes vocabulary which needs to be decided. Some extensions are required to represent the domain-specific dimensions (specimen, analyte, location).
Each record in the ESB can link directly to the information from those specialized systems. Ideally those provide a back-reference, enabling two-way navigation.

In addition to the information systems mentioned so far, there are numerous specialized systems operated independently from governmental agencies, e.g. Chemical Entities of Biological Interest ChEBI[15] or GeoNames[16]. Whether those should be referenced is merely a matter of policy - the technical opportunity exists.

Semantic Network Service (SNS)
SNS[17] has been developed since 2001 based on ISO Topic Maps[18] and the XML Topic Maps interface. Unfortunately the Topic Maps community has rejected a fusion with the Semantic Web which means we have to abandon their paradigm.

SNS includes a thesaurus, a gazetteer, and a chronicle. The thesaurus has already been implemented based on iQvoc, the Simple Knowledge Management System (SKOS) and the complementing "Extension for Lables" (SKOS-XL). The gazetteer is currently being implemented in a similar way, combining SKOS and the GeoNames Ontology. The chronicle will have to follow, based on SKOS and the Linked Events Ontology[19].

2.4 Data Lifting

Most databases at the Agency are not able to render RDF, and many of them don't even have any defined interfaces like a Web Service or CSV export. We have to take some examples and look for reference solutions for typical cases.

One example should be the library metadata system which is also used to describe research projects. This legacy system is not maintained anymore and may be replaced in the future, possibly based on an RDF representation of the data. It provides a classical OPAC interface, and this may be the key to access the data from outside.

Another example is the already mentioned GSBL, which has a Web Service interface to provide its Web client with the data, and it may provide LED as well.

Currently under development is the Soil Metadata Portal which will include an INSPIRE[20] compliant Web Catalogue Service (CSW). This year's INSPIRE conference which will take place in Istanbul at the end of June will host a tutorial on Geographical Linked Data[21], and we will carefully observe the patterns presented there, as

[15] http://www.ebi.ac.uk/chebi/
[16] http://www.geonames.org/
[17] http://www.semantic-network.de
[18] http://isotopicmaps.org/
[19] http://linkedevents.org/ontology/
[20] http://inspire.jrc.ec.europa.eu/
[21] http://datalift.org/en/node/21

implementing INSPIRE through Linked Data is not yet regulated (and in INSPIRE everything has to be regulated).

If there is absolutely no existing data interface we have to go down to the physical data model and use D2RQ[22], but most of the legacy data models are badly documented and rather cryptic.

2.5 Front End

So far we will have millions (or even billions) of HTTP URIs that can be resolved in RDF, we have links to be followed, and we have a SPARQL endpoint. This is not enough to convince humans (and especially decision makers) of any added information value – we need a human-oriented interface so they can explore the data and visualize the results in tables, diagrams, and maps. This should be generic enough to work on any data that conforms the supported standards (SKOS, SCOVO, etc.), but should also specific enough to compete with the native user interfaces of the integrated systems.

Some of these systems have very elaborate interfaces dealing with all the subtleness of their respective individual model, and we will have to leave some of this to them. We cannot go into every individual detail, but we offer a transparent integration point for all.

As the registry knows all the properties and notably which properties link between databases, it should be possible to demonstrate walk-throughs like starting with a specimen in the ESB, look-up the GSBL about the characteristics of the observed substance and then retrieve all the soil observation programs dealing with the same substance and maybe share location with the ESB specimen. This is something that has been envisioned by decision makers for many times but it never has come true.

2.6 Sustainability

In the domain of environmental protection sustainability is a strategic asset, and this should also be valid in case of information systems. Many of the systems we are talking about have been working over 10 years and more, and the outcome of LED should be able to do the same.

In parts this is an organisational matter that cannot be regulated by the LED project, but the implementation can support easy continuation and evolution.

Linked Data contributions that make data available once and then move over to the next node will not survive. So the key issue is implementing self-updating interfaces, either by direct life access to the native production data or by continuous incremental one-way synchronisation into the LED triple store.

The second key issue is the transparency of the integration work bench so that further systems can be easily integrated even after the LED project has been completed.

[22] http://d2rq.org/

3 Summary and Conclusion

The launch of a dedicated R&D project by the German agency will raise the previous LED initiatives to a new level by:

- implementing a national core cloud with links to the EEA terminology and nature information system;
- developing sustainable integration patterns and tools;
- producing reusable software components that may be adopted by others.
- establishing a comprehensive reference terminology on the national level;
- providing an intuitive user interface on top of the most convenient RDF standards (SKOS, SCOVO …);
- generating added information value by cross-database walk-through patterns.

As usual in research, we cannot anticipate the outcome in detail, and there may be unpredictable ideas at any time during the contract period. Anyway, as the data is provided by a governmental agency, LED will provide a reliable, always topical information source to the public.

References

See also Web-links in footnotes on the previous pages.

1. Berners-Lee, T.: Linked Data. W3C Design Issues. (2006/9). http://www.w3.org/DesignIssues/LinkedData.html
2. Convention on Access to Information, Public Participation in Decision-making and Access to Justice in Environmental Matters" by the United Nations Economic Commission for Europe (UNECE). http://www.unece.org/fileadmin/DAM/env/pp/documents/cep43e.pdf
3. Directive 2003/4/EC of the European Parliament and of the Council of 28 January 2003 on public access to environmental information and repealing Council Directive 90/313/EEC. http://europa.eu/legislation_summaries/environment/general_provisions/l28091_en.htm
4. Heath, T., Bizer, C.: Linked Data: Evolving the Web into a Global Data Space. (1st edition). Synthesis Lectures on the Semantic Web: Theory and Technology, 1:1, 1-136. Morgan & Claypool (2011). http://linkeddatabook.com/editions/1.0/
5. Bandholtz, T., Schleidt, K.: Summary of W4 eEnvironment Terminology. In: Hřebíček, J., Hradec, J., Pelikán, E., Mírovský, O., Pillmann, W., Holoubek, I., Bandholtz, T. (Eds.): Proceedings of the European conference of the Czech Presidency of the Council of the EU TOWARDS eENVIRONMENT. Opportunities of SEIS and SISE: Integrating Environmental Knowledge in Europe. Prague (2009) http://www.e-envi2009.org/SummaryTerminologyW4.pdf
6. Hodge, G.: Report on the Outcome of the Ecoterm V Workshop, U.N. Food and Agriculture Organization, Rome 5-6 October 2009. (2010) http://projects.eionet.europa.eu/ecoinformatics/library/ecoinformatics_indicator/ecoterm_5-6102009