Geophysical Research Abstracts
Vol. 12, EGU2010-2079, 2010
EGU General Assembly 2010

Linked Data: what does it offer Earth Sciences?

Simon Cox and Sven Schade

European Commission Joint Research Centre, Institute for Environment and Sustainability, Ispra, Italy
(simon.cox@jrc.ec.europa.eu, +39 0332 78 6325)

'Linked Data' is a current buzz-phrase promoting access to various forms of data on the internet. It starts from the two principles that have underpinned the architecture and scalability of the World Wide Web:
1. Universal Resource Identifiers – using the http protocol which is supported by the DNS system.
2. Hypertext – in which URIs of related resources are embedded within a document.

Browsing is the key mode of interaction, with traversal of links between resources under control of the client. Linked Data also adds, or re-emphasizes:
• Content negotiation – whereby the client uses http headers to tell the service what representation of a resource is acceptable,
• Semantic Web principles – formal semantics for links, following the RDF data model and encoding, and
• The 'mashup' effect – in which original and unexpected value may emerge from reuse of data, even if published in raw or unpolished form.

Linked Data promotes typed links to all kinds of data, so is where the semantic web meets the 'deep web', i.e. resources which may be accessed using web protocols, but are in representations not indexed by search engines.

Earth sciences are data rich, but with a strong legacy of specialized formats managed and processed by disconnected applications. However, most contemporary research problems require a cross-disciplinary approach, in which the heterogeneity resulting from that legacy is a significant challenge. In this context, Linked Data clearly has much to offer the earth sciences. But, there are some important questions to answer.

What is a resource? Most earth science data is organized in arrays and databases. A subset useful for a particular study is usually identified by a parameterized query. The Linked Data paradigm emerged from the world of documents, and will often only resolve data-sets. It is impractical to create even nested navigation resources containing links to all potentially useful objects or subsets. From the viewpoint of human user interfaces, the browse metaphor, which has been such an important part of the success of the web, must be augmented with other interaction mechanisms, including query.

What are the impacts on search and metadata? Hypertext provides links selected by the page provider. However, science should endeavor to be exhaustive in its use of data. Resource discovery through links must be supplemented by more systematic data discovery through search. Conversely, the crawlers that generate search indexes must be fed by resource providers (a) serving navigation pages with links to every dataset (b) adding enough 'metadata' (semantics) on each link to effectively populate the indexes. Linked Data makes this easier due to its integration with semantic web technologies, including structured vocabularies.

What is the relation between structured data and Linked Data? Linked Data has focused on web-pages (primarily HTML) for human browsing, and RDF for semantics, assuming that other representations are opaque. However, this overlooks the wealth of XML data on the web, some of which is structured according to XML Schemas that provide semantics. Technical applications can use content-negotiation to get a structured representation, and exploit its semantics. Particularly relevant for earth sciences are data representations based on OGC

Geography Markup Language (GML), such as GeoSciML, O&M and MOLES. GML was strongly influenced by RDF, and typed links are intrinsic: xlink:href plays the role that rdf:resource does in RDF representations. Services which expose GML-formatted resources (such as OGC Web Feature Service) are a prototype of Linked Data.

Giving credit where it is due. Organizations investing in data collection may be reluctant to publish the raw data prior to completing an initial analysis. To encourage early data publication the system must provide suitable incentives, and citation analysis must recognize the increasing diversity of publication routes and forms. Linked Data makes it easier to include rich citation information when data is both published and used.