

Missing(?) Building Blocks for a Linked Data Ecosystem

Axel Polleres
Siemens AG Österreich

State of LOD affairs:

- Is Linked Data the “Silver Bullet” for data integration?
 - We’ve come up with a nice interoperable data format (RDF) & a bunch of publishing principles
 - Enables lightweight data integration in a bottom-up, schema-less fashion

- LOD Cloud and table of datasets 339 datasets

**# of LOD Datasets
on CKAN**

Jun-13	339	(+44)
...		
Sep-11	295	(+92)
Sep-10	203	(+108)
Jul-09	95	

- ... slower growth rate or people too busy to let us know?
- More worrying or even harmful: Stale datasets...
 - { [] http://data.nytimes.com/elements/latest_use ?Y }
 - “2010-06-14”

Use case...

- Can we collect and use relevant open Data for a concrete domain (in our case: indicators that determine sustainability aspects of cities) in an automated fashion?

The image shows a screenshot of a Siemens website article titled "Daten-Pipeline für Stadtstaaten" (Data Pipeline for City States). The article discusses the development of a data pipeline for city data to analyze sustainability factors. The text includes: "Nachhaltigere Städte durch Offene Daten", "Siemens baut eine Daten-Pipeline für Stadtstaaten.", "Welche Faktoren bestimmen die Nachhaltigkeit von Städten? Wie verändern sich diese im Laufe der Zeit? Will man Herausforderungen wie Klimawandel, demographischen Veränderungen oder Urbanisierung gewachsen sein, braucht man Antworten auf diese Fragen.", and "Ähnlich einer Web-Suchmaschine Pipeline öffentliche Stadtstaaten vor Wikipedia und Webportalen. Ca. 2 mehr als 300 Stadtstaaten sind derzeit laufend aktualisiert und erweitert."

Below the article is a whiteboard diagram illustrating the data pipeline. The diagram shows a flow from input sources (PDF, CSV, Focus Crawler) to a central processing stage (City Data, GIS, RDF, Analysis & Modeling), which then outputs to various applications (GIS, APIs, Report Generator). The diagram is drawn in blue ink on a whiteboard.

Suspicion/Experience/Lessons learnt

- Openness may raise eyebrows... API keys, different license models, etc. Different implications on reuse, e.g.
 - republication
 - aggregation
 - How to implement pricing models? access control mechanisms? provenance tracking?
- Dynamicity vs. stale Datasets makes LOD idea hard to sell in an industry context:
 - e.g. NYT dataset (not even speaking of stream-data and the like here)
- OWL+RDFS are not enough for integration
 - Numerical data within implicit information

Missing(?) Building Blocks for a Linked Data Ecosystem:

- Not all Linked Data is open, needs mechanisms to declare & ***compose*** licenses.
- Linked Data needs standard mechanisms to deal with Dynamicity & Evolution.
- Linked Data Quality needs standard mechanisms to deal with Provenance (& Trust)
- Linked Data needs more (and less) than OWL
- and BTW: not all Linked data is RDF (JSON, CSV)...

Goal:

My expectation on the seminar is a road-map outlining

- 1. how these building blocks can implemented in terms of industry-strength standards & best practices that can interplay and scale on the Web
- 2. where further fundamental research is necessary