LOD2 Stack Tutorial

A tutorial by
Sebastian Tramp, University of Leipzig
Hugh Williams, OpenLink Software
Katja Eck, Wolters Kluwer Germany
Tutorial Presenters

Sebastian Tramp
University of Leipzig

Hugh Williams
Openlink Software

Katja Eck
Wolters Kluwer Germany
Tutorial Contents

09:30 - 09:45 Introduction
09:45 - 10:30 Virtuoso Universal Server
10:30 - 11:00 Coffee Break
11:00 - 11:20 The Semantic Data Wiki Ontowiki
11:20 - 11:45 LOD2 Toolstack - its Usage in an Industrial Environment
11:45 - 12:15 Linking with the Silk Workbench
12:15 – 12:30 Discussion and Buffer
12:30 – 14:00 Lunch Break
How to use the USB flash drive

1. Copy the complete content to your harddisk
2. Install the VirtualBox software
3. Import the Open Virtual Appliance (LOD2.ova)
4. Make a snapshot (to revert changes later)
5. Start the Virtual Machine
LOD2 STACK INTRODUCTION
Table Of Contents

• LOD2 – Why, What & When
• LOD2 stack – easy access to Linked Data software
• LOD2 stack walk through demonstration
LOD2
Why, What & When
Creating Knowledge out of Interlinked Data

From the Web of Documents to the Semantic Data Web

**Data Web** (since 2006)
- URI de-referencability
- Web Data integration
- RDF serializations

**Semantic Web** (Vision 1998, starting ???)
- Reasoning
- Logic, Rules
- Trust

**Social Web** (since 2003)
- Folksonomies/Tagging
- Reputation, sharing
- Groups, relationships

**Web** (since 1992)
- HTTP
- HTML/CSS/JavaScript

W3C Semantic Web

Web 2.0
Creating Knowledge out of Interlinked Data

What works now? What has to be done?

- Web - a global, distributed platform for data, information and knowledge integration
- exposing, sharing, and connecting pieces of data, information, and knowledge on the Semantic Web using URIs and RDF

**Achievements**

1. Extension of the Web with a **data commons** (50B facts)
2. vibrant, global RTD community
3. **Industrial uptake** begins (e.g. BBC, Thomson Reuters, Eli Lilly, WKD)
4. Emerging **governmental adoption** in sight (e.g. EC Open Data Portal)
5. Establishing Linked Data as a **deployment path** for the Semantic Web.

**Challenges**

1. **Coherence**: Relatively few, expensively maintained links
2. **Quality**: partly low quality data and inconsistencies
3. **Performance**: Still substantial penalties compared to relational
4. **Data consumption**: large-scale processing, schema mapping and data fusion still in its infancy
5. **Usability**: Establishing direct end-user tools and network effect
LOD2, a FP7 EU funded project

Creating Knowledge out of Interlinked Data

12 - DBpedia 3.7 released, including 15 localized editions

The LOD2 team is very happy to announce the release of DBpedia 3.7. The new release is based on Wikipedia dumps dating from late July 2011. The new DBpedia data set describes more than 3.64 million things, of which 1.83 million are classified in a consistent ontology, including 416,000 persons, 526,000 places, 106,000 music albums. Continue reading... read

12 - A part of the LOD2 Technology Stack: Sig.ma Enterprise Edition (EE) available

Sig.ma demonstrate a quite pragmatic component developed in the LOD2 Stack. Sig.ma EE will be distributed in the first release of the integrated version of the stack that will be released in late 2011 by the LOD2 project. The original Sig.ma service was created as a demonstration of live, on-the-fly Web of... Continue reading... read

12 - Indian-summer school on Linked Data started

Today we started in Leipzig the Indian-summer school on Linked Data. More than 30 participants are attending the summer school. We have lined up a seasoned crew of lecturers from the LOD2 consortium and beyond, including Chris Bizer and Robert Estele from LOD2 partner FUB, Jens Lehmann, Axel Igonga and Sebastian Tramp from AILW, Vojtech... Continue reading... read

LOD2 is a large-scale integrating project co-funded by the European Commission within the FP7 Information and Communication Technologies Work Programme (Grant Agreement: No. 257943). Commencing in September 2010, this 4-year project comprises leading Linked Open Data technology researchers, companies, and service providers from across 7 European countries and is coordinated by the AILW research group at the University of Leipzig.

Over the past 3 years, the semantic web activity has gained momentum with the widespread publishing of structured data as RDF. The Linked Data paradigm has therefore evolved from a practical
Objectives of LOD2

• LOD2 project objectives:
  • Increase visibility of Linked Data activities
  • Improve the software technology which support it
  • Support deployment Linked Data components
  • Improve information sharing between Linked Data components so that publishing Linked Data is eased.
  • Improve access to the content: the online Linked Open Data.

• Core enabler and end-user accessible result: the LOD2 stack
A strong partnership
LOD2 STACK

EASY ACCESS TO LINKED DATA SOFTWARE
Creating Knowledge out of Interlinked Data

The Linked Open Data Life Cycle

- Interlinking/Fusing
- Classification/Enrichment
- Quality Analysis
- Evolution/Repair
- Search/Browsing/Exploration
- Extraction
- Storage/Querying
- Manual revision/authoring
Creating Knowledge out of Interlinked Data

LOD2 stack anno 2011: easing deployment

LOD2 stack v2 = More components + more inter component integration

LOD2 stack v1 = Debian repository

2010
Sindice
Silk
Virtuoso
D2R
CKAN

2011
Sigma.EE
PoolParty
OntoWiki
Dbpedia
ORE

2012
Installing the LOD2 stack - system requirements

We standardize on Ubuntu 12.04
- Most components are ubuntu release independent.

or a Linux distribution which supports Debian packages

the software in the stack is open-source
- although individual licenses differ
- Some components are also available as commercial product
- The source itself is not (yet) distributed through the LOD2 stack repository.
Installing the LOD2 stack – software installation

### Requirements

In order to install OntoWiki from the tip (the most recently added changeset in the distributed revision control tool Mercurial) you need:

#### Mercurial

Distributed version control system version 1.6 or higher; version 1.7 or greater highly recommended.

For Ubuntu 10.04 (Lucid Lynx):

- Add new PPA and install Mercurial:

  ```
  sudo apt-add-repository ppa:mercurial-ppa/releases
  sudo apt-get update
  sudo apt-get install mercurial
  ```

- Add new PPA and update an existing Mercurial:

  ```
  sudo apt-add-repository ppa:mercurial-ppa/releases
  sudo apt-get update
  sudo apt-get upgrade
  ```

Note: If you do not have apt-get-repository, you should install it with

```
sudo apt-get install python-software-properties
```

For Mac OS X 10.4 and higher:

1. Install MacPorts
2. Depending on your preferred shell, issue one of the following commands in Terminal app:

   ```
   sudo port install mercurial -bash_completion
   ```

   or

   ```
   sudo port install mercurial +bash_completion
   ```

Note: Mercurial depends on python26 which depends on tk which in turn depends on X11. If you do not want to build X11, install python26 with the no_tkinter variant before installing Mercurial:

```
sudo port install python26 -no_tkinter
```
Linked Data publishing capabilities currently offered

- Covers most of the LOD publishing cycle
- Combination of
  - locally installed software,
  - online available software, and
  - online available data sources as well as data packages
  - about page in the LOD demonstrator (http://demo.lod2.eu/lod2demo)

Disclaimer. No harmonized user interface.
## Current list of Components

<table>
<thead>
<tr>
<th>Locally installed component</th>
<th>Online component</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Virtuoso (RDFstore, isparql, faceted browsing, sponger, ...)</strong></td>
<td>PoolParty</td>
</tr>
<tr>
<td><strong>Ontowiki</strong></td>
<td>PoolParty Extractor</td>
</tr>
<tr>
<td>owcli, LibErfurt</td>
<td>CKAN (publicdata.eu)</td>
</tr>
<tr>
<td><strong>Silk</strong></td>
<td>Sindice</td>
</tr>
<tr>
<td>Limes &amp; Colanut</td>
<td>Sigma</td>
</tr>
<tr>
<td>Valiant</td>
<td>Spotlight</td>
</tr>
<tr>
<td>Semantic Spatial Browser</td>
<td>LODstat</td>
</tr>
<tr>
<td>Sigma.EE</td>
<td></td>
</tr>
<tr>
<td>D2R (with CORDIS example)</td>
<td></td>
</tr>
<tr>
<td>ORE</td>
<td></td>
</tr>
<tr>
<td>DLlearner</td>
<td></td>
</tr>
</tbody>
</table>
Components foreseen to be added

- SPARQL editor
- General Statistical data visualizer
- R2R mapping tool (Rdf to Rdfl)
- Linking environment
- Link Data statistics
- Link Data manager: an ETL tool

We are open for new components and willing to support the integration into the stack!
LOD2 Stack – Virtuoso Universal Server

The Data Storage Layer by Hugh Williams, OpenLink Software

http://lod2.eu
Topics Covered

- **OpenLink Software**
  - About OpenLink Software
  - Core Platform behind Linked Open Data (LOD) Cloud
  - Linked Data projects snapshot

- **Product Overview & Features**
  - Why is Virtuoso Important to LOD2
  - Data Virtualization Middleware
  - Pluggable Linked Data Cartridges
  - Sophisticated Content Crawler
  - Insight Discovery & Exploration
  - Powerful SPARQL Query Service
  - Powerful SPARQL Query Builder

- **Demonstration – Loading and Querying CKAN Datasets**

- **Performance Tuning**

- **Future Enhancements**

- **Questions & Answers**
OpenLink Software

• Is a 19 year old leading provider of high-performance, scalable, and secure technology covering:
  - Data access middleware drivers/providers
  - Data virtualization middleware
  - Native database management (combined RDBMS or Graph Store)
  - Enterprise collaboration.

• Respective product portfolio offerings include:
  - UDA High-Performance drivers/providers for ODBC, JDBC, ADO.NET, XMLA
  - Virtuoso Universal Server
  - OpenLink Data Spaces for socially enhanced personal and/or enterprise collaboration.
Cont’d OpenLink Software

• A W3C Member and participant in Semantic Web Related work groups (SPARQL 1.1 & RDB2RDF etc)

• Leading supporter and participant of the original Linked Open Data (LOD) project

• One of the founding members of the DBpedia project along with Free University of Berlin & Leipzig University

• A LOD2 project consortium member
Core Platform behind Linked Open Data (LOD) Cloud

Core Platform
(Graph DBMS and Linked Data Deployment) behind DBpedia, many bubbles in the LOD Cloud, and the LOD Cloud cache itself.
### Virtuoso Linked Data projects snapshot

- **DBpedia** - public SPARQL endpoint over the DBpedia data (and [international Chapters](#))
- **LOD Cloud Cache** - public server hosting [LOD cloud](#) datasets
- **URIBurner** - [Linked Data](#) generation & transformation service
- **Linked Geo Data** - OpenStreetMap Spatial data as [Linked Data](#)
- **Sindice** - SPARQL endpoint behind its Semantic Web Index
- **Data.gov** - US Government [Linked Data](#)
- **Health.data.gov** - Clinical Quality [Linked Data](#) on [health.data.gov](#)
- **Seevl** - [Linked Data](#) *music discovery service*
- **Bio2RDF** - Life science data mapped to [Linked Data](#)
- **Neurocommons** - Life science data mapped to [Linked Data](#)
- **Musicbrainz** - MusicBrainz database published as [Linked Data](#)
- **Others** - Many others …
Virtuoso Universal Server

(Product Overview & Features)
Enterprise and Individual Agility via Data Virtualization, without compromising performance, scalability, and security.
Creating Knowledge out of Interlinked Data

Why is Virtuoso Important to LOD2

- Linked Data Deployment modulo the following challenges
  - De-referencable URI complexities
  - URI style (hash or slash) distractions
  - Loose Coupling of Information and Data
  - SPARQL endpoint commissioning
  - Linked Data Views over Relational Data (incl. R2RML support)
  - Faceted Browsing
  - Proven Performance & Scalability.
An in-built middleware layer ("Sponger") for creating Transient & Persistent Views over Heterogeneous Data Sources.
Pluggable Linked Data Cartridges

A collection of prefabricated and customizable Data Extraction, Transformation, and Lookup cartridges (drivers) covering a vast ranges of data formats and data access protocols.
Sophisticated Content Crawler

DBMS hosted Content Crawler that's leverages loosely coupled binding to the Sponge Middleware component for transformation of unstructured and semi-structured data into Linked Data.
Insight Discovery & Exploration

**Native Faceted Browsing** that enables multi-dimensional drill-downs via any browser
Powerful SPARQL Query Service

Basic SPARQL Endpoint for Creating Query Definitions & Sharing Query Results.
Powerful SPARQL Query Builder

Use Query By Example (QBE) Patterns to **Construct** & **Share** Query Results.
Demonstration

Loading CKAN LOD Datasets into Virtuoso and query as Linked Data

• This demonstration shows how the LOD2 Stack can be used for Loading CKAN Linked Open Data datasets which are part of the LOD2 Stack into the Virtuoso Quad Store resulting in the automatic deployment of the loaded datasets as Linked Data by Virtuoso enabling them to be discovered, traversed and navigated using Linked Data tools.

• Requires [Virtuoso Faceted Browser](http://lod2.eu) VAD installation
Demonstration - Loading CKAN Datasets into Virtuoso

This is Version 1.0 of the LOD2 Stack, which comprises a number of tools for managing the life-cycle of Linked Data. The life-cycle comprises in particular the stages:

- Extraction of RDF from text, XML and SQL
- Querying and Exploration using SPARQL
- Authoring of Linked Data using a Semantic Wiki
- Semi-automatic link discovery between Linked Data sources
- Knowledge-base Enrichment and Repair

You can access tools for each of these stages using the menu on top.

The LOD2 Stack is developed by the LOD2 project consortium comprising 15 research groups and companies. The LOD2 project is co-funded by the European Commission within the 7th Framework Programme (GA no. 237954).

You can find further information about the LOD2 Stack and the LOD2 project at http://lod2.eu.
Demonstration - Loading CKAN Datasets into Virtuoso

This is Version 1.0 of the LOD2 Stack, which comprises a number of tools for managing the life-cycle of Linked Data. The life-cycle comprises in particular the stages:

- Extraction of RDF from text, XML and SQL.
- Querying and Exploration using SPARQL.
- Authoring of Linked Data using a Semantic Wiki.
- Semi-automatic link discovery between Linked Data sources.
- Knowledge-base enrichment and repair.

You can access tools for each of these stages using the menu on top.

The LOD2 Stack is developed by the LOD2 project consortium comprising 13 research groups and companies. The LOD2 project is co-funded by the European Commission within the 7th Framework Programme (GA no. 257934).

You can find further information about the LOD2 Stack and the LOD2 project at http://lod2.eu.
Demonstration - Loading CKAN Datasets into Virtuoso

The LOD2 stack includes downloadable datasets as debian packages. These packages are extracted from CKAN if you select one of the links below the system will invoke the corresponding debian package update procedure. Therefore they can only be installed when the Lod2 demonstrator is used on the machine itself as it requires administration rights.

- ckan-dataset-museums-in-italy
- ckan-dataset-mondial
- ckan-dataset-linked-open-senate
- ckan-dataset-lichfield-spending
- ckan-dataset-italian-public-schools-linkedopendata-it
- ckan-dataset-lp-1166-2-data
- ckan-dataset-grr
- ckan-dataset-ch normalized-am erica
- ckan-dataset-ch-logged
Demonstration - Loading CKAN Datasets into Virtuoso
Demonstration - Loading CKAN Datasets into Virtuoso
Demonstration - Loading CKAN Datasets into Virtuoso
Demonstration - Loading CKAN Datasets into Virtuoso
Demonstration - Loading CKAN Datasets into Virtuoso
Performance Tuning

- Memory Allocation & Disk configuration
  - NumberOf Buffers
  - MaxDirtyBuffers
  - MaxCheckpointRemap
  - Disk Striping
  - Linux “swappiness”

- ServerThreads
  - HTTP allocated on Server startup
  - SQL allocated on demand for both external and internal use

- Indexes
  - Default 2 full indices over RDF quads plus 3 partial indices should suffice for most use cases
  - Custom Indexes can be created

- **Status** command for server status and **explain()** function for query plans

- Performance tuning and diagnostics documentation
Future Enhancements

• Column Store Edition
  • Column wise indexing and data compression
  • More suitable for RDF work loads providing major performance improvements
  • Dynamic Query optimization
  • Adaptive Caching of Joins

• Faceted Browser

• RDF Sponger Cartridges

• WebID
  • User creation and X.509 certificate generation
  • Authentication & authorisation (ACLs)
Questions & Answers
Thank you for your attention!

OpenLink Software Contact Information:

Web Site:  http://www.openlinksw.com
Virtuoso:  http://virtuoso.openlinksw.com
Support:   http://support.openlinksw.com
Weblog:    http://www.openlinksw.com/blog
LinkedIn:  http://www.linkedin.com/company/openlink-software
Twitter:   http://twitter.com/OpenLink
Google+:  http://plus.google.com/100570109519069333827/
Facebook: http://facebook.com/OpenLinkSoftware
LOD2 is a large-scale integrating project co-funded by the European Commission within the FP7 Information and Communication Technologies Work Programme. This 4-year project comprises leading Linked Open Data technology researchers, companies, and service providers. Coming from across 12 countries the partners are coordinated by the Agile Knowledge Engineering and Semantic Web Research Group at the University of Leipzig, Germany. LOD2 will integrate and syndicate Linked Data with existing large-scale applications. The project shows the benefits in the scenarios of Media and Publishing, Corporate Data intranets and eGovernment.
• One of the oldest (founded 1409) and
• largest (30,000 students) universities in Germany
• Institute for Applied Computer Science (InfAI)
AKSW

- Agile Knowledge Engineering and Semantic Web
- Founded in 2006
- AKSW aims:
  - Contributing to the advancement of science in Semantic Web, Knowledge Engineering, Software Engineering
  - Cost efficient, high-impact R&D, which proves usefulness at an early stage
  - Bridge the gap between research results and applications
- 25+ researchers
- Subgroups:
OntoWiki as part of the LOD2 Stack - the idea, the power & benefits

- Collection of tools to support Linked Data publication
- OntoWiki:
  - Extraction
  - Storage / Querying
  - Manual revision / authoring
  - Search / Browsing / Exploration
  - (Evolution)
OntoWiki Overview

Semantic Data Wiki
OntoWiki Overview contd.

Resources over articles
OntoWiki Overview contd.

Forms over markup
OntoWiki Vision

1. Generic data wiki for RDF models
   • No data model mismatch (structured vs. unstructured)

2. Application framework:
   • Knowledge-intensive applications,
   • Agile processes,
   • Distributed user groups
OntoWiki Features

- Knowledge Bases (aka. graphs)
  - Linked Data enabled or not
- Generic list and resource views
- Versioning
- Commenting on arbitrary resources
- User management + access control
- Inline editing
- Navigation hierarchies (e.g. Class hierarchies)
OntoWiki Features contd.

• Search
• Linked Data Server
• Linked Data Client
• Import/export of RDF/XML, Turtle, RDF/JSON
• View/import arbitrary resources
• Filtering
• SPARQL editor
OntoWiki Interfaces

- SPARQL Endpoint
- Linked Data Endpoint
- REST API
- Command Line Interface
OntoWiki Architecture

- RDFauthor
- Extensions
- OntoWiki Application Framework
- Erfurt API
- Zend Framework

http://lod2.eu
OntoWiki Extensibility

- Components, Plugins, Modules, Wrapper
- Views/Templates
- Themes
- Localizations
LOD2 Tool Stack – its Usage in an Industrial Environment

By Katja Eck, Wolters Kluwer Germany
Agenda

- Wolters Kluwer Germany Company Profile & Content Supply Chain
- Use Case Description
- Demo Application: LOD2 Stack in Use
Wolters Kluwer Germany Company Profile

Wolters Kluwer Germany (WKG):
“Semantic Technologies and Standards are an enabler for the media and publishing industry to create added-value for their customers with reasonable costs.”

WKG Legal & Regulatory

<table>
<thead>
<tr>
<th>Companies/Brands</th>
<th>Products (Examples)</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Carl Heymanns Verlag</td>
<td>- IP, Administrative Law</td>
</tr>
<tr>
<td>- Luchterhand</td>
<td>- Civil, Family, Labor Law</td>
</tr>
<tr>
<td>- Werner Verlag</td>
<td>- Construction Law</td>
</tr>
<tr>
<td>- Carl Link</td>
<td>- Publications for Schools/KiTas</td>
</tr>
<tr>
<td>- CW Haarfeld</td>
<td>- Public Health Insurance</td>
</tr>
<tr>
<td>- Deutscher Wirtschaftsdienst</td>
<td>- Magazin „Personalwirtschaft“ (HR Management)</td>
</tr>
<tr>
<td>- AnNoText</td>
<td>- SW for Lawyers and Notaries</td>
</tr>
<tr>
<td>- Jurion</td>
<td></td>
</tr>
</tbody>
</table>

WKG Tax & Accounting

<table>
<thead>
<tr>
<th>Companies/Brands</th>
<th>Products (Examples)</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Akademische Arbeits-gemeinschaft Verlag</td>
<td>- Tax SW for Consumers</td>
</tr>
<tr>
<td>- Addison Group</td>
<td>- SW for Tax Accountants</td>
</tr>
<tr>
<td>- Tsenit CS Plus</td>
<td>- SW for SMEs with focus</td>
</tr>
<tr>
<td></td>
<td>- Controlling and Accounting</td>
</tr>
</tbody>
</table>

WKG is part of Wolters Kluwer B.V.

<table>
<thead>
<tr>
<th>Customer orientation</th>
<th>Worldwide reach</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Lawyers</td>
<td>- Europe</td>
</tr>
<tr>
<td>- Tax Accountants</td>
<td>- North America</td>
</tr>
<tr>
<td>- Corporations and SMEs</td>
<td>- Asia/Pacific</td>
</tr>
<tr>
<td>- Financial institutions</td>
<td>- Economic success</td>
</tr>
<tr>
<td>- Health Providers</td>
<td>- Revenue 2011 EUR 3.3 bln.</td>
</tr>
<tr>
<td>- Public Sector</td>
<td>- 19.000 Employees</td>
</tr>
<tr>
<td></td>
<td>- Listed Amsterdam SE</td>
</tr>
</tbody>
</table>

Economic success
- Revenue 2011 EUR 3.3 bln.
- 19.000 Employees
- Listed Amsterdam SE
Current Content Supply Chain

Content Supply Chain of WKG

Content Acquisition
- Manually collecting data from different sources
- Most information is publicly not available

Content Enrichment
- Composing/Bundling
- Using internal taxonomies and thesauri
- Mainly manual enrichment
- Linking of WK content only

Composing Bundling

Publishing Interfacing
- Publishing mainly in the context of a distinct product
- Publishing of texts, not information

Sales
- Customer Interfacing
- Online libraries as isolated applications
- Hardly any integration with Web content
- Only first steps in integration of client software and content

Customer Service

Customer
WKG as a Consumer of LOD Data

Content Supply Chain of Wolters Kluwer Germany (WKG)

Content Acquisition
- Acquisition of LOD governmental data
  - Laws & Regulations
  - Court cases
  - Administrative Rulings
  - Statistical information
- Based on:
  - Adequate delivery format
  - Adequate metadata
  - Adequate Licensing and IPR

Content Enrichment
- Enrichment of WKG data
  - Enrichment with additional metadata from the LOD cloud
  - Automatic Interlinking within WKG data, but also into the LOD cloud
- Based on:
  - Adequate delivery format
  - Adequate metadata
  - Adequate functionality
  - Adequate Licensing and IPR

Enterprise Applications
- Data integration in Enterprise and other Costumer Applications
  - Integration of customer and WKG data with data from the LOD cloud
  - Development of new services, e.g. around metadata economics
- Based on:
  - Adequate functionality
  - Adequate APIs
  - Adequate Licensing and IPR
Creating Knowledge out of Interlinked Data

WKG as a Publisher of LOD Data

Content Supply Chain of
Wolters Kluwer Germany
(WKG)

Content Acquisition

Editing

Composing Bundling

Publishing Interfacing

Sales

Customer Service

Customer

Cloud - Publishing

Development of WKpedia
- Publishing of enriched governmental information
- Publishing of legal domain thesauri
- Motivating contextualisation in LOD cloud

Based on:
- Adequate functionality
- Adequate APIs
- Adequate Licensing and IPR

Marketing measures

Integration in overall marketing strategy of WKG
- Dissemination of LOD2 in media and publishing sector
- Launching surveys
- Permanent information of customers
- Sponsoring of conferences

Based on:
- Clear scope of LOD2 project to support future publishing paradigms
Use Case Description - Tasks of the Media & Publishing Use Case

- Adopting and deploying the LOD2 stack to the data sets of Wolters Kluwer
- Automatic interlinking and semantic enrichment of the Wolters Kluwer data sets
- Developing crucial functionality for publishing, searching, browsing and exploring interlinked legal information
- IPR management (licensing and management of usage rights) of interoperable metadata
- Investigation of „good characteristics“ of interoperable metadata
Creating Knowledge out of Interlinked Data

Prototypical System Architecture

[Diagram showing the system architecture with nodes and arrows indicating flow]

Third party documents → Silk Workbench

Silk Workbench → Extraction Service

Extraction Service → Metadata Management System, Index + LOD Cache

Metadata Management System, Index + LOD Cache → OntoWiki

OntoWiki → Domain Specific Application, e.g. AnnoText

Domain Specific Application, e.g. AnnoText → Semantic Meshups

Semantic Meshups → LOD Cloud

Mapping → Other Thesauri

LOD2 Stack Tutorial . 27.05.2012 . Page 8 http://lod2.eu
What has been done so far

✓ Delivery of content (around 900,000 documents)
  - Laws and Regulations
  - Court decisions
  - Books, Journals
  - … in different legal domains
✓ Implementation of technical Infrastructure
✓ Completion of Mapping Schema & Mapping Script for RDF Transformation
✓ Upload, Curation and Management of controlled vocabulary
✓ Triplification of all delivered content
✓ Mapping of generated RDF metadata with the correct concept-uri's defined in the controlled vocabularies

April 2012 - First release of the news & media data sets
Creating Knowledge out of Interlinked Data

What will be done in 2012

- Interlinking with DBpedia via classification and mapping algorithms
- Merging WK datasets with metadata from DBpedia
- Further optimization of already used toolset (usability of SILK and OntoWiki, better integration of tools, …)

December 2012 – First release of interlinked news and media metadata sets
Demo of the stack in use –
Management, Publishing & Enrichment of Court Metadata

http://lod2.eu
Keine Begründung eines „fliegenden Gerichtsstandes“ für Urheberverletzungen im Internet

Comments, Descriptions and Notes
- There are no comments, descriptions or notes on this knowledge base.

Suggestions | Outgoing references | Documents that refer to this document

Meta Data
- Date of Creation: 2006-02-04
- Language: de
- Reference: ag_frankfurt-main_31%20C%20201141%202F09_2009-08-21
- Title: Keine Begründung eines „fliegenden Gerichtsstandes“ für Urheberverletzungen im Internet

Court: AG Frankfurt/Main
Issuer: WP37
Associated Practice: Verfahrensrecht
Area: Urheberrecht, Zivilprozessrecht
Experiences

- Project as a chance to participate in the development and shape the landscape of published information in the legal domain
- Tools are working good for our purposes, e.g. 95% of the silk links are reasonable, about 75% are exact (depending on the concepts), in this early stage
- Nonetheless much work is left: functionalities, interfaces and new tools have to be adapted for our specific use cases with regard to...
  - Usability - Logic behind is difficult to understand for non technicians – therefore usability has to be improved to ease the handling of metadata
  - Workflow Processes
  - Visualization – tools have to stay manageable for large amounts of data
  - Data repair and cleansing has to be supported

-> influenced by our domain, users, data amount, products, …

- Beside the tool stack, huge effort is necessary for data transformation due to schema development and cleansing of inconsistencies
Contact

Katja Eck

Mail: keck@wolterskluwer.de
Skype: katja.eck
Tel.: +49 8936007-3127

Wolters Kluwer Deutschland GmbH
Freisinger Straße 3
85716 Unterschleißheim/ München
LOD - Creating Knowledge out of Interlinked Data
LOD2 is a large-scale integrating project co-funded by the European Commission within the FP7 Information and Communication Technologies Work Programme. This 4-year project comprises leading Linked Open Data technology researchers, companies, and service providers. Coming from across 12 countries the partners are coordinated by the Agile Knowledge Engineering and Semantic Web Research Group at the University of Leipzig, Germany.

LOD2 will integrate and syndicate Linked Data with existing large-scale applications. The project shows the benefits in the scenarios of Media and Publishing, Corporate Data intranets and eGovernment.
Web-based Systems Group

- School of Business & Economics, Freie Universität Berlin
- Research focus: Linked Data technologies for extending the World Wide Web with a global data commons
- Funded Projects:
  - LOD2 - Creating Knowledge out of Interlinked Data
  - LATC - LOD Around The Clock
  - PlanetData
- Visit us at: http://wbsg.de
The Web of Data is a single global data space because data sources are connected by links.

Over 30 billion triples published as Linked Open Data (09/19/2011).

But:
- Less than 500 million links
- Most publishers only link to one other dataset

LOD data sets by the number of other data sources that are target of outgoing RDF links.
Challenges for Link Discovery

- The Web of Data is heterogeneous
  - Many different vocabularies are in use
  - Different data formats
  - Many different ways to represent the same information

Distribution of the most widely used vocabularies
Challenges for Link Discovery

- Large range of domains
  - 277 data sources in the LOD cloud from a variety of domains
  - Linkage Rules are different in each domain
  - Writing a Linkage Rule is for each of these domains is usually not trivial

Distribution of triples by domain
Challenges for Link Discovery

• Scalability
  • The current LOD cloud contains 277 datasets (August 2011)
  • Over 31 billion triples in total
  • Infeasible to compare every possible entity pair

<table>
<thead>
<tr>
<th>Domain</th>
<th>Number of datasets</th>
<th>Triples</th>
<th>%</th>
<th>(Out-)Links</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Media</td>
<td>27</td>
<td>1,855,413,060</td>
<td>5.88</td>
<td>50,469,665</td>
<td>10.10</td>
</tr>
<tr>
<td>Geographic</td>
<td>26</td>
<td>6,111,263,253</td>
<td>19.36</td>
<td>35,751,295</td>
<td>7.16</td>
</tr>
<tr>
<td>Government</td>
<td>45</td>
<td>13,302,470,684</td>
<td>42.14</td>
<td>19,353,426</td>
<td>3.87</td>
</tr>
<tr>
<td>Publications</td>
<td>86</td>
<td>2,958,969,764</td>
<td>9.37</td>
<td>135,925,930</td>
<td>27.21</td>
</tr>
<tr>
<td>Life sciences</td>
<td>42</td>
<td>3,042,142,230</td>
<td>9.64</td>
<td>191,825,949</td>
<td>38.40</td>
</tr>
<tr>
<td>User-generated content</td>
<td>14</td>
<td>115,072,057</td>
<td>0.36</td>
<td>3,431,983</td>
<td>0.69</td>
</tr>
<tr>
<td>LOD datasets per domain</td>
<td>277</td>
<td>31,568,522,702</td>
<td>499,564,104</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

http://lod2.eu
Link Discovery Tools

- Tools enable data publishers to set links
- Most tools generate links based on user-defined linkage rules
- A linkage rule specifies the conditions data items must fulfill in order to be interlinked
- Popular Link Discover Tools:
  - Silk Link Discovery Framework
  - LIMES
  - Others: http://esw.w3.org/TaskForces/CommunityProjects/LinkingOpenData/EquivalenceMining
Silk Link Discovery Framework

- Tool for discovering links between data items within different Linked Data sources.
- The Silk Link Specification Language (Silk-LSL) allows to express complex linkage rules
- Can be used to generate owl:sameAs links as well as other relationships
- Scalability and high performance through efficient data handling
Silk Versions

- **Silk Single Machine**
  - Generate links on a single machine
  - Local or remote data sets
- **Silk MapReduce**
  - Generate RDF links using a cluster of multiple machines
  - Based on Hadoop (Can be run on Amazon Elastic MapReduce)
- **Silk Server**
  - Provides an HTTP API for matching instances from an incoming stream of RDF data while keeping track of known entities
  - Can be used as an identity resolution component within applications that consume Linked Data from the Web
(Simplified) Linking Workflow

Select Datasets
- Select two data sources
- Select the entity types to be interlinked

Write Linkage Rule
- Specifies how two entities are compared
- Can be written manually or learned

Generate Links
- Locally or on a Hadoop Cluster
- Write Links to file or a triple store
A linkage rule is represented as a tree consisting of 4 types of operators:

<table>
<thead>
<tr>
<th>RDF paths</th>
<th>Transformations</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Similar to SPARQL 1.1 Property Paths</td>
<td>• Transforms the result set of an RDF paths</td>
</tr>
<tr>
<td>• Examples:</td>
<td>• Variety of built-in transformations</td>
</tr>
<tr>
<td>• ?movie/dbpedia:director/rdfs:label</td>
<td>• Examples:</td>
</tr>
<tr>
<td>• ?person/label[@lang='en']</td>
<td>• LowerCase</td>
</tr>
<tr>
<td></td>
<td>• RegexReplace</td>
</tr>
<tr>
<td></td>
<td>• Stem</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Similarity Metrics</th>
<th>Aggregations</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Similarity of two inputs based on a user-defined metric.</td>
<td>• Aggregates multiple similarity metrics</td>
</tr>
<tr>
<td>• Examples:</td>
<td>• Examples:</td>
</tr>
<tr>
<td>• Various string similarity metrics</td>
<td>• Min, Max, Average</td>
</tr>
<tr>
<td>• Geographic similarity</td>
<td>• Quadratic Mean</td>
</tr>
<tr>
<td>• Date similarity</td>
<td>• Geometric Mean</td>
</tr>
</tbody>
</table>
Example: Interlinking cities

: Aggregation
  function: minimum

: Comparison
  distanceMeasure = Levenshtein
  threshold = 1

: Comparison
  distanceMeasure = wgs84
  threshold = 1

: Transformation
  func. = lowerCase

: Transformation
  func. = lowerCase

: Property
  property = coords

: Property
  property = coords

: Property
  property = label

: Property
  property = label
Silk Workbench

- Silk Workbench is a web application which guides the user through the process of interlinking different data sources.
- Enables the user to manage different sets of data sources and linking tasks.
- Offers a graphical editor which enables the user to easily create and edit linkage rules.
- Offers tools to evaluate the current linkage rule.
- Includes support for learning linkage rules.
Workspace

The Workspace holds a set of projects consisting of:

- **Data Sources**
  - Holds all information that is needed by Silk to retrieve entities from it.
  - Usually a file dump or a SPARQL endpoint

- **Linking Tasks**
  - Interlinks a type of entity between two data sources
  - e.g. Interlinking movies in DBpedia and LinkedMDB
Linkage Rule Editor

- Allows to view and edit linkage rules
- Linkage Rules are shown as a tree
- Editing using drag & drop.
Learning Linkage Rules

- Linkage Rules can be learned interactively
- Can be used to generate new linkage rules or to improve existing rules
- Learned Linkage Rule can be viewed and edited by the user
Availability

- Silk can be downloaded from the official homepage at:
  
  http://www4.wiwiss.fu-berlin.de/bizer/silk/

- Support is provided through the official mailing list:
  
  http://groups.google.com/group/silk-discussion

- The latest source code is available from the project's Git repository and can be browsed online at:
  
  http://www.assembla.com/code/silk/git/nodes/

- Silk is licensed under the terms of the Apache Software
Demo 1: Interlinking Movies

- Interlinking movies between two data sources:
  - **DBpedia**: Linked dataset extracted from Wikipedia
  - **LinkedMDB**: Large dataset for movies
- For demonstration, we assume that no existing links are available